Technical Report 1357

Moderators of the Tailored Adaptive Personality Assessment System Validity

Stephen Stark, Oleksandr S. Chernyshenko, Christopher D. Nye, Fritz Drasgow
Drasgow Consulting Group

Leonard A. White
U.S. Army Research Institute

July 2017

United States Army Research Institute for the Behavioral and Social Sciences

**U.S. Army Research Institute
for the Behavioral and Social Sciences**

**A Directorate of the Department of the Army
Deputy Chief of Staff, G1**

Authorized and approved:

**MICHELLE SAMS, Ph.D.
Director**

Research accomplished under contract
for the Department of Defense by

Drasgow Consulting Group

Technical Review by

J. Douglas Dressel, U.S. Army Research Institute
Irwin José, U.S. Army Research Institute

NOTICES

DISTRIBUTION: This Technical Report has been submitted to the Defense Information Technical Center (DTIC). Address correspondence concerning reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: DAPE-ARI-ZXM, 6000 6th Street (Bldg. 1464 / Mail Stop: 5610), Fort Belvoir, Virginia 22060-5610.

FINAL DISTRIBUTION: Destroy this Technical Report when it is no longer needed. Do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE *(DD-MM-YYYY)* July 2017	2. REPORT TYPE Final	3. DATES COVERED *(From - To)* September 2010 - September 2013
4. TITLE AND SUBTITLE Moderators of the Tailored Adaptive Personality Assessment System Validity		5a. CONTRACT/GRANT NUMBER GS09Q10DFD0506 9Q0SSTIS308
		5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S) Stephen Stark, Oleksandr S. Chernyshenko, Christopher D. Nye, Fritz Drasgow and Leonard A. White		5c. PROJECT NUMBER A790
		5d. TASK NUMBER 329
		5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Drasgow Consulting Group 3508 N. Highcross Rd. Urbana, IL 61802		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 6000 6th Street, Fort Belvoir, VA 22060-5610 Defense Manpower Data Center 4000 Gigling Road, Seaside, CA 93955		10. SPONSOR/MONITOR'S ACRONYM(S) ARI 11. SPONSORING/MONITORING NUMBER Technical Report 1357
12. DISTRIBUTION AVAILABILITY STATEMENT Distribution Statement A: Approved for public release; distribution is unlimited.		
13. SUPPLEMENTARY NOTES ARI Research POC: Dr. Leonard White, Personnel Assessment Research Unit		

14. ABSTRACT

The Army is conducting an evaluation of a new measure of personality, the Tailored Adaptive Personality Assessment System (TAPAS), for possible use to augment the predictive power of the Armed Services Vocational Aptitude Battery (ASVAB) for personnel selection and classification decisions. Historically, the assessment of one's personality has been plagued by inauthentic responding such as "faking good" and unmotivated responding. It is possible that some individuals do not honestly "do their best" as they answer TAPAS items or attempt to "game" TAPAS and employ a response strategy that leads to invalid and misleading scores. As a result of such aberrant responding, estimates of the validity of TAPAS for predicting important outcome variables may be biased toward zero. The present research examined the effects of such aberrant responding on the criterion-related validity of TAPAS and, in addition, evaluated whether individuals engaging in aberrant responding gained any advantage over those who responded in accordance with the test instructions. The item response theory (IRT) method developed for identifying random responding was found to be highly effective with both nonadaptive and adaptive tests and power was somewhat lower, but better than expected for detecting strategic responding in simulation studies. In addition, when statistical response flags and indices designed to detect unusually fast and patterned responding were applied to operational TAPAS data, only relatively small proportion so of examinees were flagged. When TAPAS validities were reexamined excluding those respondents, small proportions of the effects on criterion-related validities were minimal, suggesting that aberrant responding has little effect on utility.

15. SUBJECT TERMS

Personality assessment, Selection and classification, Moderator effects, Aberrance detection

SECURITY CLASSIFICATION OF:			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21.RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited Unclassified	70	Tonia S. Heffner 703-545-4408

Technical Report 1357

Moderators of the Tailored Adaptive Personality Assessment System Validity

Stephen Stark, Oleksandr S. Chernyshenko, Christopher D. Nye, Fritz Drasgow
Drasgow Consulting Group

Leonard A. White
U.S. Army Research Institute

Personnel Assessment Research Unit
Tonia S. Heffner, Chief

July 2017

ACKNOWLEDGMENTS

The authors are especially thankful to Dr. Daniel Segall, Defense Manpower Data Center, and Dr. Tonia Heffner, U.S. Army Research Institute for the Behavioral and Social Sciences, for their keen insights and suggestions in connection with this effort.

MODERATORS OF THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM VALIDITY

EXECUTIVE SUMMARY

Research Requirement:

The Army is conducting an evaluation of a new measure of personality, the Tailored Adaptive Personality Assessment System (TAPAS), for possible use to augment the predictive power of the Armed Services Vocational Aptitude Battery (ASVAB) for personnel selection and classification decisions. Historically, the assessment of one's personality has been plagued by inauthentic responding such as "faking good" and unmotivated responding. It is possible that some individuals do not honestly "do their best" as they answer TAPAS items or attempt to "game" TAPAS and employ a response strategy that leads to invalid and misleading scores. As a result of such aberrant responding, estimates of the validity of TAPAS for predicting important outcome variables may be biased toward zero. The present research examined the effects of such aberrant responding on the criterion-related validity of TAPAS and, in addition, evaluated whether individuals engaging in aberrant responding gained any advantage over those who responded in accordance with the test instructions.

Procedure:

To understand the effects of unmotivated TAPAS responding and possible ways to detect or thwart it, we conducted 3 investigations:

1. A simulation analysis where aberrant responses (random or strategic) were generated and validity decreases were calculated for different levels of aberrance. We also evaluated whether changing prior distributions of trait scores would result in lower scores for unmotivated respondents and improvements in observed validities.

2. We developed a statistical method for identifying unmotivated examinees based on the concept of appropriateness measurement (Levine & Drasgow, 1982), which is sometimes called person fit. Specifically, we adapted Drasgow, Levine and Williams's (1985) ℓ_z person fit statistic for use with multidimensional pairwise preference (MDPP) tests of any dimensionality and conducted a simulation investigation to examine the effectiveness of that index with static and adaptive personality tests.

3. We examined actual TAPAS item response data from 31,996 U.S. Army applicants, who took either a static or adaptive version of TAPAS, and flagged those who appeared to provide unusually fast, random, patterned, or strategic responses. Applicant response patterns were screened using item response times, a Markov chain statistic designed to detect patterned responding, and the new ℓ_z statistic. We then investigated the extent to which removing unmotivated respondents improved criterion validities of TAPAS scores and composites.

Findings:

Investigation 1 found that the overall validity of TAPAS dimensions was only minimally affected when random or strategic responding was present. Even in conditions where 40% of simulees responded randomly or strategically to 50% of items, validities declined by only .02-.04 relative to the values in normal responding conditions. Validity decrements in the .06-.11 range were observed only when 40% (or more) simulees responded aberrantly to all items. As expected, validity decrements were smaller overall in CAT conditions than in static conditions because of the higher test information associated with adaptive item selection. In addition, using Normal (-1,1) and Beta (3,7) priors did not seem to improve observed validities, but, as expected, did substantially increase the bias in estimated trait scores. Thus, changing the N(0,1) scoring priors currently implemented in TAPAS would likely have only small effects on criterion validities.

In Investigation 2, we developed a statistical method for identifying atypical response patterns by adapting Drasgow, Levine, and Williams' (1985) ℓ_z person fit statistic for use with multidimensional pairwise preference (MDPP) tests. A Monte Carlo investigation found that the MDPP ℓ_z index had nearly perfect power to detecting 100% random responding and power at or above .90 for detecting strategic responding with static tests, even with critical values corresponding to a nominal alpha (Type I error rate) of .01. In addition, when just 50% of the simulated responses were aberrant, power was .84 or higher for random responding and .60 or higher for strategic responding at the .01 alpha level. As expected, power was somewhat lower, but still good, for adaptive tests, with values of .87 or above for detecting 100% random responding and .59 or above for detecting 100% strategic responding based on critical values corresponding to an alpha of .05.

Finally, we examined actual TAPAS item response data from 31,996 U.S. Army applicants, who took either a static or adaptive version, and flagged those who appeared to provide unusually fast, random, patterned, or strategic responses. Only relatively small percentages of examinees were flagged. When TAPAS validities were reexamined excluding those respondents, the effects on criterion validities were minimal, suggesting that aberrant responding has had little effect on utility.

Utilization and Dissemination of Findings:

The methods developed in this research were found to be effective for identifying examinees who provide unusually fast, random, patterned, or strategic responding. Even when such responses were present, TAPAS was found to provide scores with little diminution in criterion related validity and therefore can enhance enlistment screening.

MODERATORS OF THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM VALIDITY

CONTENTS

Page

LIST OF TABLES

Page

Page

LIST OF FIGURES

MODERATORS OF THE TAILORED ADAPTIVE PERSONALITY ASSESSMENT SYSTEM VALIDITY

CHAPTER 1: INTRODUCTION

In collaboration with the Army, Military Entrance Processing Command (MEPCOM), and Defense Manpower Data Center (DMDC), Drasgow Consulting Group (DCG) has been evaluating the Tailored Adaptive Personality Assessment System (TAPAS) for use in personnel screening and classification contexts. Army applicants have taken TAPAS at Military Enlistment Processing Stations (MEPS) since May of 2009. Although TAPAS "counts," in the sense that some Army applicants have been screened out based on low scores, the number is very small and it is possible that some individuals do not "do their best." Unmotivated responding or other attempts to “game the system” could produce misleading scores and have detrimental effects on validities for predicting outcomes such as attrition, adjustment to Army life, disciplinary incidents, and overall performance.

Concerns about unmotivated or, more generally, aberrant responding are not limited to the personality assessment domain. In the context of cognitive ability testing, for example, individuals may be inadequately measured because they cheat on some test items, respond randomly, misgrid some items on a paper-and-pencil test (e.g., answering item 21 in the space provided for item 20), have atypical educations, and so forth. Schmitt, Chan, Sacco, McFarland, and Jennings (1999) showed that the test-criterion validity coefficients can be greatly reduced when such individuals are present.

Understanding the effects of aberrant responding on test validities is important for several reasons. First, random, careless or “fake good” responding can obscure the true relationships between personality variables and important criteria and lead to the erroneous conclusion that such variables are unimportant in selection contexts. Second, aberrant responding can affect test norms as well as the choice of cut off scores, so, at the very least, examinees exhibiting aberrant response patterns should be flagged prior to establishing norms and conducting validity analyses. Finally, research into the reliable identification of aberrant responders can inform the development of policies and strategies to encourage motivated, accurate responding, which can be beneficial to both the organization and the examinees.

Purpose of this Research

To understand the effects of unmotivated TAPAS responding and possible ways to detect or thwart it, we conducted three investigations:

1. A simulation investigation where aberrant responses (random or strategic) were generated and validity decreases were calculated for different levels of aberrance. We also evaluated whether changing prior distributions of trait scores would result in lower scores for unmotivated respondents and improvements in observed validities.

2. We developed a statistical method for identifying unmotivated examinees based on the concept of appropriateness measurement (Levine & Drasgow, 1982), which is sometimes called person fit. Specifically, we adapted Drasgow et al.’s ℓ_z person fit statistic for use

with multidimensional pairwise preference (MDPP) tests of any dimensionality and conducted a simulation investigation to examine the effectiveness of that index with static and adaptive personality tests.

3. We examined actual TAPAS item response data from 31,996 U.S. Army applicants, who took either a static or adaptive version of TAPAS, and flagged those who appeared to provide unusually fast, random, patterned, or strategic responses. Applicant response patterns were screened using item response times, a Markov chain statistic designed to detect patterned responding, and the new ℓ_z statistic. We then investigated the extent to which removing unmotivated respondents improved criterion validities of TAPAS scores and composites.

Before we delve into the details of these investigations, we briefly describe the TAPAS assessments that have been used in the MEPS, present the trait definitions, and review the item response theory (IRT) methods underlying test scoring.

Description of TAPAS testing at MEPS

In collaboration with the Army Research Institute (ARI) for the Behavioral and Social Sciences, DCG has developed a series of computerized forms of TAPAS for MEPS testing. These forms utilized a pool of over 800 personality statements capable of generating thousands of pairwise preference items. Statement parameters for this pool were estimated from data collected in large samples of new recruits from 2006 to 2008 (Drasgow, Stark, Chernyshenko, Nye, Hulin, & White, 2012).

The first TAPAS form was a 104-item, 13-dimension (13D) computerized adaptive test (CAT). This form was administered from May 4, 2009 to July 10, 2009 to about 2,200 Army and Air Force recruits. In July 2009, TAPAS MEPS testing was expanded to 15 dimensions and test length was increased to 120 pairwise preference items. One form of this 15D test was static, meaning that all examinees answered the same sequence of items. We refer to this form throughout this report as the TAPAS-15D-Static. The other form was adaptive; each examinee received items tailored to his or her trait level estimates. This form is referred to herein as the TAPAS-15D-CAT.

TAPAS-15D-Static was administered to all eligible examinees from mid-July to mid-September of 2009 and then phased out. TAPAS-15D-CAT was introduced in September of 2009 and testing continued until July of 2011 when the original statement pool was replaced with a new item pool.

Table 1 below shows the 15 narrow personality dimensions (facets) assessed by TAPAS-15D-Static and TAPAS-15D-CAT. As illustrated, all of the Big Five personality factors (Goldberg, 1990) are represented by two or more facets. Specifically, the forms measured four facets of Conscientiousness (Self-Control, Achievement, Order, and Non-Delinquency), three facets of Extraversion (Dominance, Sociability, Attention Seeking), three facets of Emotional Stability (Adjustment, Optimism, and Even Tempered), two facets of Openness to Experience (Intellectual Efficiency and Tolerance), and two facets of Agreeableness (Cooperation and

Consideration). Physical Conditioning was also measured, because it predicts many important outcomes in military contexts (Drasgow, Stark, Chernyshenko, Nye, Hulin, & White, 2012).

Table 1. Dimensions Assessed by TAPAS-15D-Static and TAPAS-15D-CAT

TAPAS Facet Name	Brief Description	"Big Five" Broad Factor
Dominance	High scoring individuals are domineering, "take charge" and are often referred to by their peers as "natural leaders."	Extraversion
Sociability	High scoring individuals tend to seek out and initiate social interactions.	
Attention Seeking	High scoring individuals tend to engage in behaviors that attract social attention; they are loud, loquacious, entertaining, and even boastful.	
Generosity	High scoring individuals are generous with their time and resources.	Agreeableness
Cooperation	High scoring individuals are trusting, cordial, non-critical, and easy to get along with.	
Achievement	High scoring individuals are seen as hard working, ambitious, confident, and resourceful.	Conscientiousness
Order	High scoring individuals tend to organize tasks and activities and desire to maintain neat and clean surroundings.	
Self-Control	High scoring individuals tend to be cautious, levelheaded, able to delay gratification, and patient.	
Non-Delinquency	High scoring individuals tend to comply with rules, customs, norms, and expectations, and they tend not to challenge authority.	
Adjustment	High scoring individuals are worry free, and handle stress well; low scoring individuals are generally high strung, self-conscious and apprehensive.	Emotional Stability
Even Tempered	High scoring individuals tend to be calm and stable. They don't often exhibit anger, hostility, or aggression.	
Optimism	High scoring individuals have a positive outlook on life and tend to experience joy and a sense of well-being.	
Intellectual Efficiency	High scoring individuals are able to process information quickly and would be described by others as knowledgeable, astute, and intellectual.	Openness To Experience
Tolerance	High scoring individuals are interested in other cultures and opinions that may differ from their own. They are willing to adapt to novel environments and situations.	
Physical Conditioning	High scoring individuals tend to engage in activities to maintain their physical fitness and are more likely to participate in vigorous sports or exercise.	Military-Specific

The administration procedures for TAPAS-15D-Static and TAPAS-15D-CAT were identical. Each testing session was initiated by a MEPCOM test administrator who entered the examinee's social security number into the computer. Next, each examinee was asked to read information related to the purpose of the assessment and instructions about answering TAPAS items. Testing then began and continued until all items were completed or the 30-minute time limit elapsed. Detailed results for each TAPAS testing session were saved and transferred to a central database. These data included trait scores, the number of seconds taken to complete the test, flags to detect fast responders, relevant item response data, and scores on two selection composites referred to as Can Do and Will Do. These composites were developed using a large sample of Army trainees to predict basic training outcomes as well as Army life adjustment, support for peers and physical fitness (see Knapp & Heffner, 2010). TAPAS scores were considered "valid" only if an examinee completed at least 80% of the items. (Note that in the event of a test interruption, the administrator could save the session and restart the assessment at the same point).

TAPAS Scoring

TAPAS scoring is based on the Multi-Unidimensional Pairwise Preference (MUPP) IRT model originally proposed by Stark (2002). The model assumes that when a respondent encounters stimuli *s* and *t* (which, in this case, correspond to two personality statements), the respondent considers whether to endorse *s* and, independently, considers whether to endorse *t*. This process of independently considering the two stimuli continues until one and only one stimulus is endorsed. A preference judgment can then be represented by the joint outcome (Agree with *s*, Disagree with *t*) or (Disagree with *s*, Agree with *t*). Using a 1 to indicate agreement and a 0 to indicate disagreement, the outcome (1,0) indicates that statement *s* was endorsed but statement *t* was not, leading to the decision that *s* was preferred to statement *t*; an outcome of (0,1) indicates that stimulus *t* was preferred to *s*. Thus, the probability of endorsing a stimulus *s* over a stimulus *t* can be formally written as

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_{st}\{1,0 \mid \theta_{d_s}, \theta_{d_t}\}}{P_{st}\{1,0 \mid \theta_{d_s}, \theta_{d_t}\} + P_{st}\{0,1 \mid \theta_{d_s}, \theta_{d_t}\}},$$

where:

$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) =$ probability of a respondent preferring statement *s* to statement *t* in item *i*,
$i =$ index for items (i.e., pairings), where $i = 1$ to I,
$d =$ index for dimensions, where $d = 1, \ldots, D$, d_s represents the dimension assessed by statement *s*, and d_t represents the dimension assessed by statement t,
$s, t =$ indices for first and second statements, respectively, in an item,
$(\theta_{d_s}, \theta_{d_t}) =$ latent trait scores for the respondent on dimensions d_s and d_t respectively,
$P_{st}(1,0 \mid \theta_{d_s}, \theta_{d_t}) =$ joint probability of endorsing stimulus *s* and not endorsing stimulus *t* given latent trait scores $(\theta_{d_s}, \theta_{d_t})$, and
$P_{st}(0,1 \mid \theta_{d_s}, \theta_{d_t}) =$ joint probability of not endorsing stimulus *s* and endorsing stimulus *t* given latent trait scores $(\theta_{d_s}, \theta_{d_t})$.

With the assumption that the two statements are evaluated independently, and with the usual IRT assumption that only θ_{d_s} influences responses to statements on dimension d_s and only θ_{d_t} influences responses to dimension d_t (i.e., local independence), we have

$$P_{(s>t)_i}(\theta_{d_s}, \theta_{d_t}) = \frac{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t})}{P_s(1 \mid \theta_{d_s}) P_t(0 \mid \theta_{d_t}) + P_s(0 \mid \theta_{d_s}) P_t(1 \mid \theta_{d_t})},$$

where

$P_s(1 \mid \theta_{d_s}), P_s(0 \mid \theta_{d_s}) =$ probability of endorsing/not endorsing stimulus *s* given the latent trait value θ_{d_s}, and

$P_t(0 \mid \theta_{d_t}), P_t(1 \mid \theta_{d_t}) =$ probability of endorsing/not endorsing stimulus *t* given latent trait θ_{d_t}.

The probability of preferring a particular statement in a pair thus depends on θ_{d_s} and θ_{d_t}, as well as the model chosen to characterize the process for responding to the individual statements. Toward that end, Stark (2002) proposed using the dichotomous case of the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000), which has been shown to fit personality data reasonably well (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Stark, Chernyshenko, Drasgow, & Williams, 2006).

TAPAS scoring is done via Bayes modal estimation (see Drasgow et al., 2012; Stark, Chernyshenko, & Drasgow, 2005). Standard errors for TAPAS trait scores are estimated using a replication method developed by Stark and colleagues (Stark, Chernyshenko, Drasgow, & White, 2012). In brief, this method involves using the IRT parameters for the items that were administered to generate 50 new response patterns based on an examinee's TAPAS trait scores. The resulting simulated response patterns are then scored and the standard deviations of the respective trait estimates over the 50 replications are used as standard errors for the original TAPAS values. As shown by Stark et al. (2012), this replication method provides standard error estimates that are much closer to the empirical (true) standard deviations than previously used approaches (i.e., based on the approximated inverse Hessian matrix or a jack-knife procedure).

CHAPTER 2: DOES CHANGING THE PRIOR CORRECT SCORES FOR RANDOM AND STRATEGIC RESPONDING AND MAINTAIN TAPAS'S CRITERION-RELATED VALIDITY?

Objective

The first step in gaining an understanding of how aberrant responding may affect TAPAS scoring and validities was to conduct a simulation investigation where the extent of aberrant responding was known. We focused on two types of aberrance: random and strategic (fake good) responding. *Random responding* occurs when examinees choose an answer without carefully considering the stimuli composing a pairwise preference item. In addition, because the TAPAS software randomizes the order of statements composing each pairwise preference item, patterned responding (e.g., repeatedly selecting option A or alternating between options A and B) also falls within this category of aberrance. *Strategic* or *fake good* responding occurs when an examinee selects an option based on its perceived social desirability rather than how closely it describes his or her typical thoughts, feelings, or actions. In other words, rather than following test instructions to choose the statement in each pairwise preference item that is "more like me," examinees repeatedly choose options they believe will increase their chances of qualifying for enlistment.

To see the extent to which aberrant responding might affect the accuracy and validity of TAPAS scores, we conducted a Monte Carlo investigation. In each condition, 1,000 simulated examinees (simulees) took a 15D, 120-item static or adaptive (CAT) MDPP test twice. In the control conditions, examinees responded according to the underlying MUPP IRT model on both occasions. In the experimental conditions, examinees responded according to the model on the first occasion and with a designated proportion of random or strategic responding on the second. For each occasion, we then computed the correlations between estimated and known trait scores, as well as the correlations with an external criterion, to see how much accuracy and validity decreased as a function of the type and degree of aberrance. In addition, we explored whether the use of prior distributions other than multivariate standard normal would result in lower scores for aberrant respondents and improvements in observed validities.

Investigation Design

First, we computed the correlations among TAPAS dimensions using data from the 15D, 120-item CAT that was being administered at the MEPS. We then assigned to each of these nominal dimensions a validity coefficient ranging from .00 to .45. This correlation matrix was subsequently used to generate 15 standard normal trait scores plus one standard normal criterion score for each of 1,000 simulees via a FORTRAN computer program. The resulting means, standard deviations and intercorrelations for the simulated test and criterion scores are shown in Table 2. As shown in the table, the means and standard deviations for the 1,000 sets of 15 trait scores, labeled 1-15, were very close to the expected values of 0 and 1, respectively, and the resulting criterion correlations, shown in the last row, were in the desired .0 to .4 range.

Table 2. Descriptive Statistics and Intercorrelations for Simulated TAPAS and Criterion Scores

Dimension	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-0.09	0.98															
2	0.01	0.99	.13														
3	-0.01	1.02	.17	.19													
4	-0.05	1.00	.44	.13	.03												
5	0.02	1.00	.17	.27	.31	-.04											
6	0.00	1.03	-.04	-.09	-.07	-.24	.05										
7	0.05	1.01	.14	.00	.32	.06	.17	.08									
8	-0.04	0.99	.27	.23	.05	.28	.12	-.09	.07								
9	0.01	1.01	.21	-.04	.22	-.05	.23	.18	.15	-.06							
10	0.02	1.02	.17	-.11	.01	.08	.00	.11	.07	.05	.06						
11	-0.03	0.99	.18	.07	-.04	.17	-.08	-.18	-.06	.06	-.04	.04					
12	0.00	0.99	.27	.08	.16	.12	.26	.18	.10	.21	.31	.19	-.09				
13	0.01	1.02	.02	.13	.26	.28	.07	-.45	.11	.01	-.10	.00	.13	-.12			
14	0.04	1.03	.15	.05	.23	.12	.17	-.06	.41	.14	.03	.04	-.07	.08	.17		
15	-0.04	1.02	.24	.31	.20	.23	.25	-.21	.04	.14	.12	.00	.11	.12	.28	.11	
Criterion	-0.05	0.98	.36	.31	.01	.28	.25	.20	.04	.42	.09	.21	.37	.10	.05	.01	.32

Note: N = 1,000 simulees.

The scores described above served as the true values for a Monte Carlo investigation that examined the effects of aberrant responding on TAPAS scoring accuracy and validity.

Five independent variables were manipulated:

a. Test type (Static, CAT);
b. Aberrance type (None, Random, Strategic);
c. Percentage of examinees responding aberrantly (20%, 40%);
d. Percentage of items that could be answered aberrantly by aberrant examinees (25%, 50%, 100%);
e. Prior type (N(0, 1), N(-1,1), Beta(3, 7).

Because independent variable c was nested within b and d was nested within c, there were a total of 78 experimental conditions, as shown in Table 3.

In each condition, simulees were administered a 15D, 120-item test having the same design specifications as TAPAS-15D-CAT and the same 1,000 true trait scores (thetas) were used to generate pairwise preference item responses. In the None conditions (see Column 2 under Type of Aberrance), every simulee answered test items according to a "normal" (MUPP IRT) model. In the Random and Strategic conditions, specified percentages of simulees (see % Ab Persons in Column 3) answered designated percentages of items (see % Ab Items in Column 4) according to an "aberrant" model.

To simulate random responding, an aberrant simulee answered a specified proportion of items randomly. For example, in the 25% Ab Items conditions, aberrant simulees provided random responses to 24 of the 120 pairwise preference items. In the 50% Ab Items conditions, aberrant simulees answered 60 of the 120 items randomly and in the 100% Ab Items conditions, answers to all items were generated randomly.

To simulate strategic responding, an aberrant simulee was given an "opportunity" to fake good on a designated subset of items. Specifically, if the *perceived* social desirability parameters of the two statements composing a designated item differed by more than 1.0 units, an aberrant simulee selected the statement in the pair with the higher perceived desirability; otherwise the simulee responded according to the normal (MUPP) model. Note that the perceived social desirabilities for these comparison judgments were obtained by sampling values from independent normal distributions, having standard deviations of 0.5, and means equal to the respective social desirability parameters in the TAPAS statement pool. For example, suppose the first statement in a pairwise preference item had a TAPAS social desirability parameter of 1.7 and the second statement had a social desirability parameter of 2.9. Perceived social desirability values would be sampled from independent normal distributions with means of 1.7 and 2.9, respectively, and a common standard deviation of 0.5. If the perceived desirabilities differed by more than 1.0, the simulee would then select the statement with the higher perceived value. The advantage of simulating strategic (fake good) responding in this way is that it incorporates random error into the judgments of social desirability similar to Thurstone's (1927) ideas about discriminal processes in paired comparison judgments. The disadvantage, however, is that this approach relies on an arbitrary decision about how big a difference in perceived desirability is needed to trigger a strategic response. If the required difference is set too high, the effects of strategic responding could be underestimated.

The simulated item response patterns were scored using one of three prior distributions (see Column 5). N(0,1) represents the independent standard normal priors implemented in TAPAS-15D-CAT. N(-1,1) represents independent normal priors having means and variances of -1 and 1, respectively. B(3,7) represents four-parameter beta priors having means of -1.2 and variances of 0.8. Whereas the N(-1,1) prior was chosen merely to lower the mean of the posterior trait score distribution, the B(3,7) prior was intended also to introduce a substantial positive skew. The purpose of exploring the alternative prior distributions was to see whether they would result in lower trait scores primarily for simulees designated as “aberrant.”

To provide benchmarks for trait score stability, accuracy and validity with and without aberrant responding, we simulated the administration of a second test in which simulees answered according to the “normal” (MUPP IRT) model. Scores from the two test administrations in the None conditions were correlated to assess test-retest reliability. Scores from the repeated test administrations in the other conditions were used to see how random and strategic responding affected trait score accuracy and validity and whether using the alternative priors mitigated those effects.

Table 3. Investigation 1 Simulation Conditions

Condition #	Static Test			
	Type of Aberrance	% Ab Persons	% Ab Items	Prior
1	None	-	-	N(0,1)
2		-	-	N(-1,1)
3		-	-	B(3,7)
4	Random	20%	20%	N(0,1)
5				N(-1,1)
6				B(3,7)
7			50%	N(0,1)
8				N(-1,1)
9				B(3,7)
10			100%	N(0,1)
11				N(-1,1)
12				B(3,7)
13		40%	20%	N(0,1)
14				N(-1,1)
15				B(3,7)
16			50%	N(0,1)
17				N(-1,1)
18				B(3,7)
19			100%	N(0,1)
20				N(-1,1)
21				B(3,7)
22	Strategic	20%	20%	N(0,1)
23				N(-1,1)
24				B(3,7)
25			50%	N(0,1)
26				N(-1,1)
27				B(3,7)
28			100%	N(0,1)
29				N(-1,1)
30				B(3,7)
31		40%	20%	N(0,1)
32				N(-1,1)
33				B(3,7)
34			50%	N(0,1)
35				N(-1,1)
36				B(3,7)
37			100%	N(0,1)
38				N(-1,1)
39				B(3,7)

Table 3 (continued)

Condition #	CAT			
	Type of Aberrance	% Ab Persons	% Ab Items	Prior
40	None	-	-	N(0,1)
41		-	-	N(-1,1)
42		-	-	B(3,7)
43	Random	20%	20%	N(0,1)
44				N(-1,1)
45				B(3,7)
46			50%	N(0,1)
47				N(-1,1)
48				B(3,7)
49			100%	N(0,1)
50				N(-1,1)
51				B(3,7)
52		40%	20%	N(0,1)
53				N(-1,1)
54				B(3,7)
55			50%	N(0,1)
56				N(-1,1)
57				B(3,7)
58			100%	N(0,1)
59				N(-1,1)
60				B(3,7)
61	Strategic	20%	20%	N(0,1)
62				N(-1,1)
63				B(3,7)
64			50%	N(0,1)
65				N(-1,1)
66				B(3,7)
67			100%	N(0,1)
68				N(-1,1)
69				B(3,7)
70		40%	20%	N(0,1)
71				N(-1,1)
72				B(3,7)
73			50%	N(0,1)
74				N(-1,1)
75				B(3,7)
76			100%	N(0,1)
77				N(-1,1)
78				B(3,7)

Results

Static Test Simulation Conditions. Table 4 presents simulation results for the static test conditions. For both parsimony and ease of interpretation, the results presented in the table were obtained by averaging across the 15 TAPAS dimensions. In Columns 5 through 8 of the table, we show various trait score accuracy statistics, such as the correlation between estimated (T) and generating (G) trait scores (Column 5), standard errors (Column 6), bias (T-G) and the average absolute error (Columns 7 and 8). Column 9 shows criterion validities computed by averaging correlations between estimated trait scores and criterion scores (Y). The last column of Table 4 shows the change in validity for each aberrant condition relative to the None, N(0,1) condition.

As expected, the best accuracy and validity results were observed when all simulees responded in accordance with the MUPP model and trait scores were estimated using the correct N(0,1) prior (see Row 1 in Table 4). The average correlation between the estimated and generating trait scores for that condition was .83 and the average standard error was .42. Both values were similar to those observed in past MDPP simulation studies involving static tests of similar length (see Stark et al., 2012). Bias and absolute bias results for this condition were also consistent with previous research. Averaging bias statistics across dimensions and over simulees produced a value near zero, but the average absolute error (0.44) indicates that there was a moderate regression to the mean effect. The average criterion validity for this condition (.099) serves as a benchmark for comparing static test validities when aberrance was present and/or alternative scoring priors were used.

Comparing the results in Rows 2 and 3 to Row 1 indicates that changing the scoring prior had the expected detrimental effects on scoring accuracy, even though all simulees responded in accordance with the MUPP model. Not only did the bias and average absolute error increase when the alternative priors were used, but the average correlation between estimated and generating trait scores and the criterion validities were somewhat affected. For the N(-1,1) prior, Avg CorrTG decreased by .02 and the Validity Change was -.012, relative to the N(0,1). For the B(3,7) prior, Avg CorrTG decreased by .04 and the Validity Change was -.013.

Examination of accuracy and validity results for the aberrant conditions reveals that, although the accuracy of estimated traits scores was negatively affected when random or strategic responding was present, observed criterion validities were only modestly affected. Even in conditions where 40% of the simulees responded randomly to 50% of the items, the average validity declined by only .02-.03 relative to the values in the normal responding conditions. Validity decrements in the .04-.06 range were observed only when 40% of the simulees responded aberrantly to all items. The same pattern of results was observed in the strategic responding conditions, although the corresponding decreases in validities were smaller. In all aberrant conditions, the B(3,7) prior resulted in the largest validity decrease.

Comparisons of results for different types of aberrance reveals that random responding had larger negative effects on the accuracy and validity of TAPAS scores than those observed in comparable strategic responding conditions. For example, in Random, 20% Ab Persons, 100% Ab Items conditions, the average corrTGs were .69, .67, and .65 for the three priors, whereas the corresponding Strategic conditions had average corrTGs of .89, .79, and .77, respectively. The poorer recovery of true trait scores in these Random conditions translated into larger average validity changes of -.017, -.026, and -.043; much smaller declines of -.006, -.011, and -.02 were observed when similar levels of Strategic responding were simulated.

One reason for such differences between comparable random and strategic conditions was the relatively low occurrence of strategic responding. Post hoc analyses of response patterns for designated strategic responders indicated that aberrant responding occurred on less than the designated percentages of aberrant items. More specifically, because of the fairly tight matching constraints in the TAPAS item selection algorithm and the perceived social desirability difference that was required to trigger a fake good response, only about a quarter of the designated aberrant items elicited a strategic response. Hence, even though aberrant simulees were given opportunities to fake on 25%, 50%, or 100% of items, the actual proportions of fake good responses were much lower in each of those conditions.

Table 4. Results for Static Test Conditions Averaged Across Dimensions

Type of Aberrance	% Ab Persons	% Ab Items	Prior	Avg CorrTG	Avg SE	Avg (T-G)	Avg |T-G|	Avg CorrTY	Validity Change
None	-	-	N(0,1)	.83	.42	-.01	.44	**.099**	-
	-	-	N(-1,1)	.81	.42	-.63	.70	.087	-.012
	-	-	B(3,7)	.79	.39	-.88	.91	.086	-.013
Random	20%	20%	N(0,1)	.80	.42	-.01	.47	.094	-.004
			N(-1,1)	.79	.42	-.64	.72	.090	-.009
			B(3,7)	.77	.39	-.89	.92	.075	-.024
		50%	N(0,1)	.77	.42	-.03	.49	.086	-.012
			N(-1,1)	.75	.42	-.65	.75	.087	-.012
			B(3,7)	.74	.39	-.91	.96	.078	-.021
		100%	N(0,1)	.69	.42	-.03	.54	.082	-.017
			N(-1,1)	.67	.42	-.67	.80	.073	-.026
			B(3,7)	.65	.39	-.93	.99	.056	-.043
	40%	20%	N(0,1)	.78	.42	-.02	.49	.090	-.009
			N(-1,1)	.77	.42	-.65	.74	.090	-.009
			B(3,7)	.75	.39	-.90	.95	.074	-.025
		50%	N(0,1)	.71	.42	-.03	.55	.078	-.021
			N(-1,1)	.69	.42	-.67	.80	.091	-.008
			B(3,7)	.68	.39	-.93	1.00	.064	-.035
		100%	N(0,1)	.56	.42	-.05	.65	.069	-.030
			N(-1,1)	.52	.42	-.71	.90	.062	-.037
			B(3,7)	.53	.39	-.97	1.08	.045	-.054
Strategic	20%	20%	N(0,1)	.82	.42	-.01	.45	.096	-.003
			N(-1,1)	.80	.42	-.63	.71	.082	-.017
			B(3,7)	.79	.39	-.88	.91	.088	-.011
		50%	N(0,1)	.81	.42	.00	.46	.096	-.003
			N(-1,1)	.80	.42	-.63	.71	.087	-.012
			B(3,7)	.78	.39	-.88	.91	.078	-.021
		100%	N(0,1)	.81	.42	.00	.47	.092	-.006
			N(-1,1)	.79	.42	-.63	.71	.087	-.011
			B(3,7)	.77	.39	-.88	.92	.079	-.020
	40%	20%	N(0,1)	.82	.42	.00	.45	.095	-.004
			N(-1,1)	.80	.42	-.62	.71	.095	-.004
			B(3,7)	.78	.39	-.88	.92	.082	-.016
		50%	N(0,1)	.81	.42	.01	.47	.096	-.002
			N(-1,1)	.78	.42	-.62	.71	.085	-.014
			B(3,7)	.77	.39	-.89	.93	.073	-.026
		100%	N(0,1)	.78	.42	.03	.49	.089	-.010
			N(-1,1)	.77	.42	-.63	.73	.091	-.008
			B(3,7)	.74	.40	-.88	.93	.077	-.022

Note: Average CorrTG = average correlations between estimated and generated trait scores across 15 TAPAS dimensions; Avg SE = average standard error of trait scores over the 15 traits and 1000 simulees; Avg (T-G) = bias in estimated trait scores; Avg |T-G| = average absolute error in estimated trait scores; Avg CorrTY = average criterion validity; Validity Change = validity decrement relative to the None, N(0,1) condition.

Because this investigation utilized a repeated measures design where each simulee was tested twice, two sets of 15 trait scores were available for each simulee. Examinees that were designated as normal responded according to the normal model on both occasions, whereas examinees that were designated *a priori* as aberrant responded according to the aberrant model on the first occasion and according to the normal model on the second. For normal simulees, the differences between the sets of scores on occasions 1 and 2 stem only from random error, which on average should be (and was verified to be) zero. However, for simulees designated as aberrant, the differences between their respective scores on occasions 1 and 2 reflect both random error and aberrance. Because the expected value of random error is zero, comparing the average scores for aberrant examinees on the two occasions reveals the overall systematic effect of aberrance.

Table 5 presents the differences between trait scores on occasions 1 (aberrant) and 2 (normal), averaged across aberrant simulees, for the 40% Ab Persons conditions. As can be seen from the table, difference scores for the aberrant simulees in Random conditions were generally negative meaning that, on average, they received lower trait scores when responding aberrantly than when responding normally. As the percentage of aberrant items increased, trait scores for aberrant simulees decreased further, but the net effects were relatively small. Even when simulees responded to 100% of items randomly, the average decrease in the average estimated trait score was only -0.11 in the N(0,1) condition, -0.22 in the N(-1,1) condition, and -0.25 in the B(3,7) condition. Thus, the use of alternative priors reduced scores for random responders relative to the comparable N(0,1) conditions, but unfortunately the average net affects were small.

Examination of results for specific dimensions indicated that scores did not decrease uniformly across the 15 dimensions. Large decreases for aberrant test administrations were observed for dimensions 3, 7, 8, 14 and 15 in the 100% Ab Items conditions, but other dimensions showed much smaller decreases, and, in the case of dimension 6, there was a large score increase for aberrant examinees.

Results for Strategic conditions indicated that, on average, faking good had little effect on static test scores. While scores for some dimensions appeared to increase, scores for the other dimensions in those same conditions decreased. Even when simulees were given an opportunity to respond strategically on all items in the static test, the overall degree of score inflation was negligible. Average score differences across the two test administrations were the smallest for the B(3,7) prior, but as in the Random conditions, the net effects of the alternative priors were small.

Table 5. Estimated Trait Score Differences between "Aberrant" and "Normal" Static Test Administrations for Aberrant Simulees in 40% Ab Persons Conditions

Type of Aberrance	% Ab Items	Prior	Dimension															Average
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Random	20%	N(0,1)	.01	-.05	-.17	-.09	.05	.19	-.12	-.04	.00	.08	.01	-.01	.04	-.14	-.09	**-.02**
		N(-1,1)	-.07	-.05	-.13	-.03	.01	.13	-.18	-.09	-.06	-.04	-.03	.06	-.04	-.20	-.06	**-.05**
		B(3,7)	-.02	-.06	-.11	-.09	-.03	.13	-.19	-.13	-.07	-.05	-.13	-.02	-.03	-.16	-.10	**-.07**
	50%	N(0,1)	-.07	-.05	-.19	-.05	.15	.46	-.29	-.19	-.04	.08	.02	.02	-.02	-.33	-.17	**-.04**
		N(-1,1)	-.13	-.04	-.41	-.12	.02	.30	-.30	-.15	-.08	-.04	.00	-.05	.02	-.45	-.26	**-.11**
		B(3,7)	-.14	-.12	-.36	-.16	-.02	.29	-.33	-.18	-.16	-.15	-.05	-.06	-.02	-.40	-.36	**-.15**
	100%	N(0,1)	-.17	-.03	-.50	-.19	.13	.71	-.55	-.27	.05	.15	.04	.04	-.01	-.77	-.32	**-.11**
		N(-1,1)	-.22	-.19	-.81	-.26	.05	.64	-.61	-.30	-.19	-.03	-.04	-.08	-.07	-.72	-.52	**-.22**
		B(3,7)	-.18	-.15	-.63	-.32	-.02	.62	-.64	-.41	-.15	-.17	-.12	-.05	-.16	-.85	-.51	**-.25**
Strategic	20%	N(0,1)	.02	.04	.00	.04	.08	-.02	.01	-.01	-.03	.06	.01	.07	-.01	.03	.08	**.02**
		N(-1,1)	-.02	-.05	-.03	.00	.01	.01	-.04	-.03	-.02	.01	.00	.07	.00	-.03	.02	**-.01**
		B(3,7)	-.01	.01	-.17	-.03	.02	-.05	-.05	-.04	-.03	.01	.05	.01	.03	.00	.00	**-.02**
	50%	N(0,1)	.08	-.04	.01	-.04	.04	.04	.00	-.02	-.03	.11	.17	.12	.00	-.01	.09	**.04**
		N(-1,1)	.07	-.01	-.10	-.05	.06	.06	-.03	-.03	-.06	.03	.09	.06	.02	.04	.10	**.02**
		B(3,7)	.02	-.03	-.14	-.10	.05	-.04	-.10	-.04	-.02	.02	.10	.11	.00	-.04	.04	**-.01**
	100%	N(0,1)	.13	.07	-.07	-.07	.20	.11	.02	.04	-.06	.18	.30	.18	.07	-.01	.20	**.09**
		N(-1,1)	.04	-.04	-.28	-.13	.15	-.06	-.06	-.09	-.06	.09	.18	.18	.01	-.02	.06	**.00**
		B(3,7)	.04	-.05	-.30	-.15	.07	.00	-.10	-.11	-.07	.07	.25	.20	.05	-.06	.06	**-.01**

Note: Sample size in each cell was 400 simulees.

.

CAT Simulation Conditions. Table 6 presents scoring accuracy and validity results for the CAT simulation conditions. As expected, trait score estimation improved substantially with adaptive item selection. The Avg CorrTG for the None, N(0,1) condition was .89 and the Avg SE was .32, despite the tests being fairly short relative to the number of dimensions assessed. Consistent with these improvements in estimation accuracy, the Avg CorrTYs also improved.

Overall, CAT results mimicked the pattern observed for static tests. In Random conditions, the validity changes were generally less than .02 unless 100% of the items were answered randomly. In Strategic conditions, validity declines were even smaller due, in part, to the low proportions of items that were actually answered aberrantly. As with the static tests, larger absolute errors (Avg |T-G|) were observed when scoring with the N(-1,1) and B(3,7) priors, and the largest validity decreases were observed when using B(3,7). Also, as expected, the Avg CorrTY values were higher for CAT than the corresponding Static conditions.

Table 6. Results for CAT Conditions Averaged Across Dimensions

Type of Aberrance	% Ab Persons	% Ab Items	Prior	Avg CorrTG	Avg SE	Avg (T-G)	Avg |T-G|	Avg CorrTY	Validity Change
None	-	-	N(0,1)	.89	.32	.01	.35	.120	-
	-	-	N(-1,1)	.88	.34	-.48	.54	.122	.002
	-	-	B(3,7)	.87	.32	-.69	.72	.107	-.013
Random	20%	25%	N(0,1)	.88	.32	.01	.38	.122	.002
			N(-1,1)	.87	.34	-.47	.56	.112	-.008
			B(3,7)	.86	.33	-.68	.72	.101	-.019
		50%	N(0,1)	.85	.32	.02	.41	.118	-.002
			N(-1,1)	.83	.34	-.48	.58	.103	-.017
			B(3,7)	.83	.33	-.69	.74	.107	-.013
		100%	N(0,1)	.78	.32	.00	.45	.100	-.020
			N(-1,1)	.77	.34	-.48	.62	.097	-.023
			B(3,7)	.76	.33	-.71	.79	.094	-.026
	40%	25%	N(0,1)	.86	.32	.01	.40	.123	.003
			N(-1,1)	.85	.34	-.48	.57	.124	.004
			B(3,7)	.84	.33	-.70	.74	.107	-.013
		50%	N(0,1)	.80	.32	.00	.46	.107	-.013
			N(-1,1)	.79	.34	-.48	.62	.113	-.007
			B(3,7)	.77	.33	-.70	.78	.104	-.016
		100%	N(0,1)	.65	.32	-.02	.55	.089	-.031
			N(-1,1)	.64	.34	-.46	.70	.078	-.042
			B(3,7)	.62	.33	-.70	.85	.051	-.069
Strategic	20%	25%	N(0,1)	.89	.32	.02	.36	.126	.006
			N(-1,1)	.88	.34	-.48	.55	.119	-.001
			B(3,7)	.87	.32	-.69	.71	.110	-.011
		50%	N(0,1)	.89	.32	.03	.37	.115	-.005
			N(-1,1)	.87	.34	-.47	.55	.109	-.011
			B(3,7)	.86	.33	-.67	.71	.100	-.020
		100%	N(0,1)	.87	.32	.05	.38	.123	.003
			N(-1,1)	.86	.34	-.45	.55	.097	-.023
			B(3,7)	.85	.33	-.66	.70	.105	-.015
	40%	25%	N(0,1)	.89	.32	.03	.36	.119	-.001
			N(-1,1)	.88	.34	-.47	.54	.115	-.005
			B(3,7)	.87	.32	-.67	.71	.112	-.008
		50%	N(0,1)	.88	.32	.05	.38	.120	.000
			N(-1,1)	.86	.34	-.45	.54	.122	.002
			B(3,7)	.84	.33	-.66	.71	.111	-.009
		100%	N(0,1)	.85	.32	.07	.41	.118	-.002
			N(-1,1)	.83	.34	-.42	.56	.104	-.016
			B(3,7)	.83	.33	-.63	.70	.091	-.029

Note: Average CorrTG = average correlations between estimated and generated trait scores across 15 TAPAS dimensions; Avg SE = average standard error of trait estimates; Avg (T-G) = average bias in estimated trait scores; Avg |T-G| = average absolute error in estimated trait scores; Avg CorrTY = average criterion validity; Validity Change = validity difference for Aberrant and None, N(0,1) conditions.

Table 7 presents the differences between trait scores on occasions 1 (aberrant) and 2 (normal), averaged across aberrant simulees, for the 40% Ab Persons CAT conditions. In contrast to the results for the Static Random conditions where many large score changes were observed, the score changes for CAT were relatively small for all 15 TAPAS dimensions and there was no clear pattern of score increase or decline. Whereas changes as large as -0.85 and +0.71 were observed in the Static conditions, the largest changes with CAT were just -0.23 and +0.22. Also unlike the Static conditions where larger overall score decreases were observed for N(-1,1) and B(3,7) priors, the largest decreases in scores were observed in the N(0,1) conditions, but the net effects were small.

Results for Strategic conditions indicated that, on average, faking good had positive but small effects on scores. As expected, scores for individual dimensions increased more when there were higher designated percentages of aberrant items, but the average scores across all 15 dimensions increased by just 0.14 to 0.16 for conditions where simulees had an opportunity to fake on all items (100% Ab Items). And, as before, the alternative priors had little effect on score differences.

Table 7. Estimated Trait Score Differences between "Aberrant" and "Normal" CAT Administrations for Aberrant Simulees in 40% Ab Persons Conditions

Type of Aberrance	% Ab Items	Prior	Dimension															Average
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Random	25%	N(0,1)	-.06	-.03	-.07	-.01	.03	.00	-.04	-.04	-.01	-.02	-.03	.01	-.01	-.10	-.04	**-.03**
		N(-1,1)	-.02	.00	-.02	.00	.01	.04	-.05	.04	.00	-.01	.01	.01	.05	-.01	-.06	**.00**
		B(3,7)	.03	.02	-.02	.00	.01	.03	-.03	.01	-.09	-.05	-.01	-.02	.01	-.02	-.01	**-.01**
	50%	N(0,1)	-.03	-.01	-.14	-.02	-.03	.03	-.07	-.03	-.03	-.06	-.07	-.03	-.05	-.16	-.05	**-.05**
		N(-1,1)	-.01	.05	-.07	.05	.04	.11	-.05	-.02	-.04	-.02	-.01	.07	.06	-.09	-.03	**.00**
		B(3,7)	-.01	-.02	-.10	-.01	.02	.11	-.06	-.03	-.15	-.02	-.02	.01	.02	-.16	-.08	**-.03**
	100%	N(0,1)	-.09	-.06	-.17	.02	-.07	.09	-.14	-.06	-.06	-.13	-.06	-.11	-.08	-.14	-.11	**-.08**
		N(-1,1)	.07	.17	-.09	.08	.12	.22	-.04	.08	.00	.04	.07	.01	-.01	-.11	.08	**.05**
		B(3,7)	-.02	.07	-.13	.03	-.05	.18	-.10	-.04	-.21	-.08	-.03	-.08	-.03	-.23	-.03	**-.05**
Strategic	25%	N(0,1)	.07	.05	.01	.04	.08	.05	-.03	.06	.02	.02	.02	.01	.01	.01	.08	**.03**
		N(-1,1)	.04	.01	.02	.04	.02	-.02	.01	.05	.01	.02	.01	.03	.01	.06	.04	**.03**
		B(3,7)	.04	.04	.05	-.02	.02	-.03	-.01	.00	-.04	.05	.07	.06	.01	-.01	.03	**.02**
	50%	N(0,1)	.09	.05	.03	.06	.09	.06	-.05	.08	.02	.06	.13	.06	.07	.01	.11	**.06**
		N(-1,1)	.09	.08	.05	.11	.14	.01	.02	.15	.04	.03	.09	.13	.12	.06	.15	**.08**
		B(3,7)	.09	.13	.00	.11	.11	.01	.02	.08	.02	.03	.11	.03	.04	.00	.14	**.06**
	100%	N(0,1)	.16	.19	.01	.17	.23	.06	.07	.22	.13	.09	.20	.19	.24	.12	.29	**.16**
		N(-1,1)	.21	.22	.03	.18	.24	.00	.06	.22	.09	.06	.29	.22	.28	.07	.22	**.16**
		B(3,7)	.17	.20	.05	.16	.21	-.01	.09	.16	.12	.06	.22	.16	.23	.08	.24	**.14**

Note: Sample size in each cell was 400 simulees.

.

Summary and Conclusions

Investigation 1 examined the effects of random and strategic responding on the accuracy and validity of TAPAS scores across increasing levels of aberrance. Six levels of random and strategic responding were simulated with the lowest level being 20% Ab Persons, 25% Ab Items and the highest being 40% Ab Persons, 100 % Ab Items. Aberrance effects were studied for static as well as adaptive tests, and simulees were scored with either N(0,1), N(-1,1), or B(3,7) priors. The repeated measures design, where simulees took each test under aberrant and normal responding conditions, allowed the systematic effects of aberrance on TAPAS scoring to be examined.

The results of the simulations provided good news and bad news. For the good news, it was found that criterion validities were only minimally affected by random or strategic responding. Even in conditions where 40% of simulees responded aberrantly to 50% of items, the average validities declined by only .02-.03 relative to the corresponding values in the normal conditions. Only when 40% of simulees were designated to respond aberrantly to all items did validities change appreciably. As expected, validities were better in CAT conditions than in static conditions across all types and levels of aberrance, because of the higher test information associated with adaptive item selection. Thus, it appears that the MUPP approach to measurement underlying TAPAS is reasonably resistant to random and strategic responding.

On the other hand, the bad news is that the N(-1,1) and B(3,7) priors did not outperform the N(0,1) prior. We had hypothesized that random responding would not cause trait estimates to rise much above the mean of the priors whereas normal responding by simulees drawn from a N(0,1) distribution would lead to higher trait scores. Thus, we expected to see fairly large mean differences across the simulated random response patterns and the normal N(0,1) simulees. For the static test condition, we did see small decrements in mean test scores for the N(-1,1) and B(3,7) priors, but these differences were not large enough to be useful in practice. For the CAT conditions, the differences were even smaller. Thus, changing the N(0,1) scoring priors currently implemented in TAPAS is unlikely to have a substantial impact on selection decisions.

The alternative to correcting trait estimates for random and strategic responding is identifying such aberrant response patterns by a class of methods that have been called “person fit” or “appropriateness measurement” (Levine & Rubin, 1979). Person fit measures quantify the degree of departure of an observed response pattern from what is expected from a normal respondent. Individuals with response patterns exceeding some cut score are classified as aberrant. Simulation studies (e.g., Drasgow, Levine, & Williams, 1985) have shown that person fit measures can provide powerful, but not perfect, identification of aberrant response patterns.

In addition to the simulations examining the consequences of changing the N(0,1) prior to either N(-1,1) and B(3,7), we conducted two additional investigations. In the next chapter, we describe the development and evaluation of a measure that is an extension of Drasgow et al.’s (1985) ℓ_z appropriateness index. This index has been found to provide powerful detection of aberrance for static tests, but for CAT it has been found to be much less effective. However, the CAT research has largely been in the context of a single test; with TAPAS, there are effectively 15 tests. For static tests, Drasgow, Levine, and McLaughlin (1991) found that multi-test appropriateness indices were substantially more powerful than single test measures. Thus, in the context of TAPAS, there is reason to be optimistic about the effectiveness of an extension of ℓ_z. After evaluating the new measure with simulation data in Chapter 3, we examined its performance with real TAPAS in Chapter 4. In addition, two other approaches to identifying aberrant data were examined in that chapter.

CHAPTER 3: DEVELOPMENT AND EVALUATION OF A PERSON FIT INDEX FOR MULTIDIMENSIONAL PAIRWISE PREFERENCE TESTS

Objective

The main objective of this investigation was to develop a person fit index that could be used to screen TAPAS MDPP test data. Toward that end, we adapted Drasgow, Levine and Williams's (1985) ℓ_z appropriateness index, which has been shown to be effective for detecting aberrance with static cognitive ability tests. To increase the chances of detecting aberrance with CAT data, which has historically been problematic, we developed a method for computing person-specific critical values that could be used for classifying TAPAS response patterns as normal or aberrant with desired false positive rates. We then conducted a simulation to investigate the effectiveness of the new index for detecting various levels of random and strategic responding.

Development of the ℓ_z Index for MDPP Tests

A test or assessment may not provide accurate measurement of some individuals even when it is reliable and valid for an overall group. For example, scores can be spuriously high for individuals who copy some answers from a more talented neighbor or come to the exam with item knowledge. Analogously, "faking good" may lead to spuriously high scores. On the other hand, scores can be "spuriously low" when, for example, language difficulties detract from an individual's performance on a math test, an examinee misgrids his or her answer sheet, and an optical scanner misreads an answer sheet. The goal of appropriateness measurement is to identify such mismeasured test takers.

A large number of person fit indices have been introduced and evaluated. For example, Meijer and Sijtsma (2001) reviewed over forty measures and Karabatsos (2003) conducted a Monte Carlo evaluation of 36 indices. One of the fit statistics that has consistently been found to perform well in such comparisons for static tests (e.g., Drasgow, Levine, & McLaughlin, 1987) is ℓ_z, which was originally introduced by Drasgow et al. (1985) as the approximately standardized log likelihood of a response pattern.

The Unidimensional ℓ_z Index. This index has been studied largely in the context of the two- and three-parameter logistic (3PL) IRT models. The 3PL model can be written

$$P(u_i = 1 \mid \theta) = c_i + \frac{1 - c_i}{1 + \exp(-1.7 a_i (\theta - b_i))},$$

where $P(u_i = 1 \mid \theta)$ is the probability of a positive or correct response to the i^{th} item for an individual with standing θ on the latent trait, a_i, b_i and c_i are the discrimination, difficulty, and lower asymptote parameters for this item, and 1.7 is a scaling constant often used for historical reasons. The log likelihood of an individual's responses can be written

$$\ell_0 = \sum_{i=1}^{n} u_i \log P(u_i = 1 \mid \hat{\theta}) + (1 - u_i) \log(1 - P(u_i = 1 \mid \hat{\theta}))$$

where $\hat{\theta}$ is an estimate of θ . The approximate expectation of this log likelihood is

$$E(\ell_0) \approx \sum_{i=1}^{n} P(u_i = 1 | \hat{\theta}) \log P_i(u_i = 1 | \hat{\theta}) + [1 - P_i(u_i = 1 | \hat{\theta})] \log[1 - P_i(u_i = 1 | \hat{\theta})].$$

The approximate variance is

$$Var(\ell_0) \approx \sum_{i=1}^{n} P_i(u_i = 1 | \hat{\theta})[1 - P_i(u_i = 1 | \hat{\theta})]\{\log \frac{P_i(u_i = 1 | \hat{\theta})}{[1 - P_i(u_i = 1 | \hat{\theta})]}\}^2.$$

Finally, the approximately standardized index is

$$\ell_z = \frac{\ell_0 - E(\ell_0)}{\sqrt{Var(\ell_0)}}.$$

In the original research (Drasgow et al., 1985) as well as in subsequent examinations (e.g., Molenaar & Hoijtink, 1990), it has been found that ℓ_z is approximately, but not exactly, standardized.

Multidimensional Extension of the ℓ_z Index. Drasgow et al. (1991) considered the case of a multi-unidimensional test battery consisting of M tests, each of which is unidimensional. In this case there are *M* latent traits, $\underset{\sim}{\theta} = (\theta_1, \theta_2, ..., \theta_M)'$ and the log likelihood of the test battery can be written

$$\ell_0 = \sum_{m=1}^{M} \sum_{i=1}^{n} u_{mi} \log P(u_{mi} = 1 | \hat{\theta}_m) + (1 - u_{mi}) \log(1 - P(u_{mi} = 1 | \hat{\theta}_m))$$

where u_{mi} is the i^{th} response on the m^{th} test and $\hat{\theta}_m$ is the ability estimate for the m^{th} test. Due to local independence, the approximate expectation and variance of the multidimensional ℓ_0 are

$$E(\ell_0) \approx \sum_{m=1}^{M} E(\ell_{m0})$$

and

$$Var(\ell_0) \approx \sum_{m=1}^{M} Var(\ell_{m0}),$$

where ℓ_{m0} denotes the log likelihood of the m^{th} test. The multidimensional ℓ_0 can then be standardized as before.

The MDPP Adaptation of ℓ_z. Drasgow et al.'s (1991) multi-test extension of ℓ_z provides a way of identifying persons who respond aberrantly on a sequence of unidimensional tests. However, with MDPP assessments, such as TAPAS, multiple dimensions are assessed simultaneously. Thus, using the notation for MDPP tests described in Chapter 1, the log likelihood for a MDPP test involving *d* dimensions and *i* items can be written as

$$\ell_0(\hat{\underset{\sim}{\theta}}) = \sum_{i=1}^{I} u_i \log P_{(s>t)_i} + (1-u_i)\log(1-P_{(s>t)_i},$$

where $\hat{\underset{\sim}{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_D)'$ is a vector of latent trait estimates, u_i is a dichotomously scored response to the i^{th} pairwise preference item, and $P_{(s>t)_i}$ is the probability that a respondent prefers statement *s* to statement *t* in item *i* given his or her trait scores $(\hat{\theta}_{d_s}, \hat{\theta}_{d_t})$ on the dimensions assessed by the item. Accordingly, the expectation and variance of variance of ℓ_0 and can be written in a manner similar to the unidimensional case above and the computation of ℓ_z is identical:

$$E(\ell_0) \approx \sum_{i=1}^{I} P_{(s>t)_i} \log P_{(s>t)_i} + (1-P_{(s>t)_i})\log(1-P_{(s>t)_i})$$

and

$$Var(\ell_0) \approx \sum_{i=1}^{I} P_{(s>t)_i}(1-P_{(s>t)_i})\left\{\log\frac{P_{(s>t)_i}}{(1-P_{(s>t)_i})}\right\}^2.$$

The Distribution of ℓ_z and Critical Values for Data Screening. The ℓ_z indices developed by Drasgow et al. focus on identifying persons who respond inconsistently with model predictions. Responding in a way that is incongruent with one's true trait scores over the course of a long test leads to large negative ℓ_z values. Thus, based on early research showing that the distribution of ℓ_z is approximately standard normal for long tests (e.g., 80 items), critical values for a one-tailed z test could be used to classify response patterns as normal or aberrant. For example, if one wants to screen response patterns with a 5% false positive rate (i.e., 5% of normal response patterns will be misclassified as aberrant), the critical ℓ_z for a lower one-tailed z test would be -1.65. If a respondent's observed ℓ_z were less than the critical value, then the response pattern would be flagged as aberrant; otherwise the pattern would be considered normal.

Over the last two decades, research on the distribution of ℓ_z with much shorter tests has shown that the distribution is typically not standard normal (e.g., Meijer & Nering, 1997; Molenar & Hoijtink, 1990; Nering, 1995) and the substitution of trait estimates for true thetas in the ℓ_z calculations leads to conservative Type I error rates and therefore lower power to detect misfit (Snijders, 2001; van Krimpen-Stroop & Meijer, 1999). CAT creates additional complexities because each examinee receives a unique set of items that are targeted to his or her trait estimate at every point during a test. Because items are selected based on item information, and information is highest when the probability of a correct answer is near 0.5, it is difficult to detect inconsistencies with respect to model predictions. This problem is exacerbated as testing progresses, because the range of item difficulties (locations) in a CAT becomes restricted as $\hat{\theta}$ converges to $\boldsymbol{\theta}$ (van Krimpen-Stroop & Meijer, 1999).

Nering (1997) examined the distribution of ℓ_z with CAT and found that the critical value of -1.65 led to conservative classification of aberrant response patterns, and he explored an alternative simulation approach to determining the distribution of ℓ_z that would account for error in trait score estimates. van Krimpen-Stroop and Meijer (1999) also explored the distribution of ℓ_z with static tests and CATs for large numbers of simulees at various levels of $\boldsymbol{\theta}$ and found that the distribution was negatively skewed across all trait levels and test lengths when calculations were based on $\hat{\theta}$ values. In addition, they found that the ℓ_z means were too large (0.5) and the variances were too small (0.6) for CAT, which they attributed to small $E(\ell_0)$ values arising from response probabilities near 0.5 due to adaptive item selection. They investigated simulating a sampling distribution of ℓ_z for every examinee, based on $\hat{\theta}$, that would approximate the empirical distribution. They found that the process worked well for static tests, but yielded too few ℓ_z scores in the lower tail for CAT.

The ideas presented in the van Krimpen-Stroop and Meijer papers stimulated our thoughts about the use of person-specific critical values for the MDPP adaptation of ℓ_z. Specifically, in an effort to address the issues of possible nonnormality, error in latent trait estimates, variations in the distribution of ℓ_z across trait levels, and the uniqueness of tests created dynamically by adaptive item selection, the following procedure was proposed for ℓ_z screening of TAPAS data.

Upon the completion of a test (whether static or adaptive), ℓ_z is computed for an examinee's response pattern using the appropriate statement parameters and final trait score estimates. Next a sampling distribution of ℓ_z is obtained for that examinee by simulating 100 normal response patterns (a unique seed is used on each replication) according to the MUPP model using the same item parameters and trait score estimates. ℓ_z is computed for each of the simulated patterns, the values are ranked in ascending order, and the values corresponding to, say, the 1st, 5th, 10th percentiles are chosen as critical values for classifying the examinee's observed response pattern as normal or aberrant with a .01, .05, or .10 false positive rate, respectively. More specifically, if the examinee's ℓ_z is less than the critical ℓ_z, then the pattern is considered aberrant; otherwise the pattern is considered normal. The efficacy of this approach was explored via the simulation investigation described below.

A Monte Carlo Investigation of the Effectiveness of the MDPP $\boldsymbol{\ell_z}$ Adaptation

To evaluate the effectiveness of the new ℓ_z index for detecting random and strategic responding, we conducted a simulation investigation involving five independent variables:

1. Test type (Static, CAT);
2. Type of aberrance (None, Random, Strategic);
3. % of aberrant examinees (0%, 100%) (nested in type of aberrance);
4. % of items that can be answered aberrantly by aberrant examinees (100%, 50%, 25%);
5. Prior type: N(0, 1), N(-1,1), Beta (p=3, q=7).

Because independent variable 3 was nested within 2 and 4 was nested within 3, there were a total of 42 experimental conditions, as shown in Table 8. In each condition, we simulated 1,000 15D, 120-item TAPAS test administrations using the same generating (true) trait scores as in Investigation 1.

In the None conditions (cells 1 through 3 and 22 through 24 shown in Table 8), all response patterns were simulated according to the MUPP model. These conditions were used to examine Type I error, defined as the proportion of normal response patterns that were misclassified as aberrant (i.e., false positives). In the remaining cells, all response patterns were designated as aberrant. These cells were used to compute power, defined as the proportion of aberrant simulees correctly identified as aberrant (i.e., hits).

Random responding was simulated in the same manner as in Investigation 1. Prior to administering a test to a simulee, a randomly chosen subset of items was designated as "aberrant." Any remaining items were designated as "normal." When presented with a normal item, the simulee responded according to the MUPP model. When presented with an aberrant item, a random response was generated by sampling a random number from a uniform distribution and assigning a 1 if the value was greater than 0.5 and assigning a 0 otherwise.

Strategic responding was also simulated in a manner similar to Investigation 1, but rather than requiring a perceived social desirability difference of 1 or greater to trigger a fake good response, the threshold was set at 0.00. That ensured a fake good response would be "attempted" on every item designated as aberrant but did not guarantee that the statement with the higher true social desirability would be selected because the standard deviations of the discriminal dispersions for perceived desirabilities remained at 0.5.

Table 8. Investigation 2 Simulation Design

	Static					CAT			
Cell #	Type of Aberrance	% Ab Persons	% Ab Items	Prior	Cell #	Type of Aberrance	% Ab Persons	% Ab Items	Prior
1	None	-	-	N(0,1)	22	None	-	-	N(0,1)
2		-	-	N(-1,1)	23		-	-	N(-1,1)
3		-	-	B(3,7)	24		-	-	B(3,7)
4	Random	100%	100%	N(0,1)	25	Random	100%	100%	N(0,1)
5				N(-1,1)	26				N(-1,1)
6				B(3,7)	27				B(3,7)
7			50%	N(0,1)	28			50%	N(0,1)
8				N(-1,1)	29				N(-1,1)
9				B(3,7)	30				B(3,7)
10			25%	N(0,1)	31			25%	N(0,1)
11				N(-1,1)	32				N(-1,1)
12				B(3,7)	33				B(3,7)
13	Strategic	100%	100%	N(0,1)	34	Strategic	100%	100%	N(0,1)
14				N(-1,1)	35				N(-1,1)
15				B(3,7)	36				B(3,7)
16			50%	N(0,1)	37			50%	N(0,1)
17				N(-1,1)	38				N(-1,1)
18				B(3,7)	39				B(3,7)
19			25%	N(0,1)	40			25%	N(0,1)
20				N(-1,1)	41				N(-1,1)
21				B(3,7)	42				B(3,7)

Results

Table 9 presents Type I error results for Investigation 2. The numerical values across the top row of the table, .01, .05, .10,… , .50, are the nominal alpha levels; i.e., the expected proportions of normal examinees that would be misclassified as aberrant using the empirical person-specific critical values. The values in the body of the table are the observed Type I error rates. As can be seen from the table, the MDPP adaptation of ℓ_z provided excellent performance for static tests with the N(0,1) priors, they were slightly inflated for the N(-1,1) priors, and considerably inflated for the B(3,7) priors. For the CAT conditions, the Type I errors were also inflated for the B(3,7) priors, but quite conservative for both normal priors.

Table 9. Type I Error Rates for MDPP ℓ_z Aberrance Detection

Cell	Test	Prior	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
1		N(0,1)	.01	.05	.10	.13	.18	.23	.28	.34	.39	.44	.48
2	Static	N(-1,1)	.02	.06	.12	.17	.22	.26	.31	.35	.41	.47	.53
3		B(3,7)	.02	.10	.18	.26	.32	.38	.43	.48	.53	.58	.63
22		N(0,1)	.00	.01	.03	.06	.09	.11	.14	.20	.24	.29	.34
23	CAT	N(-1,1)	.01	.03	.08	.12	.16	.20	.25	.30	.36	.41	.46
24		B(3,7)	.06	.15	.21	.27	.33	.40	.44	.49	.55	.60	.65

Table 10 presents the power results for ℓ_z aberrance detection. The power values in the body of the table indicate the observed proportions of aberrant simulees who were correctly classified as aberrant at five nominal alpha levels (.01, .05, .10, .25, .50). Thus, higher values indicate better performance.

As expected, power to detect random and strategic responding was higher when larger percentages of items were simulated as aberrant and, of course, power increased as the nominal alphas increased. For static tests, with alpha =.05, very high power was observed for detecting both random and strategic responding when 50% or more items were aberrant. In the 25% Ab Items conditions, power remained fairly high for detecting random responses, but dropped off considerably for detecting strategic responding.

Examination of some specific results in Table 10 indicated that ℓ_z provided surprisingly high power to detect random and strategic responding. For a nominal alpha of .05, almost perfect power was observed for Static tests in the 100% Ab Items, N(0,1) conditions, and power for CAT was just slightly lower – with values of .87 in the Random, 100% Ab Items, N(0,1) condition and .74 in the Strategic, 100% Ab Items, N(0,1) condition. Power dropped into the .55 - .80 range for the 50% Ab Items conditions, but overall the hit rates were still quite good considering the negative findings for aberrance detection with CAT in previous studies.

For the 25% Ab Items conditions, power for CAT was fairly low relative to Static tests, CAT power was just .28 in the Random, N(0,1) and .24 in the Strategic, N(0,1) conditions. The good news is that, as was shown in Investigation 1, such low levels of aberrance have minimal effects on trait score accuracy and criterion validity.

Note also that although power appeared to be higher across the board with B(3,7) priors, that result was due to inflated Type I error (see Table 9). Power was fairly similar for the N(0,1) and N(-1,1) priors, with just minor differences that can be attributed to more conservative classification with the N(0,1) scoring.

Table 10. Power Rates for MDPP ℓ_z Aberrance Detection

Test	Type of Aberrance	% Ab Items	Prior	ℓ_z Critical Value .01	.05	.10	.25	.50
Static	Random	100%	N(0,1)	.99	1.00	1.00	1.00	1.00
			N(-1,1)	1.00	1.00	1.00	1.00	1.00
			B(3,7)	1.00	1.00	1.00	1.00	1.00
		50%	N(0,1)	.84	.96	.98	1.00	1.00
			N(-1,1)	.87	.97	.99	1.00	1.00
			B(3,7)	.91	.99	1.00	1.00	1.00
		25%	N(0,1)	.37	.66	.77	.91	.97
			N(-1,1)	.41	.70	.82	.93	.99
			B(3,7)	.49	.75	.83	.94	.98
	Strategic	100%	N(0,1)	.90	.99	1.00	1.00	1.00
			N(-1,1)	.93	.99	1.00	1.00	1.00
			B(3,7)	.97	1.00	1.00	1.00	1.00
		50%	N(0,1)	.53	.79	.89	.96	1.00
			N(-1,1)	.60	.84	.91	.97	.99
			B(3,7)	.67	.89	.95	.99	1.00
		25%	N(0,1)	.19	.44	.57	.78	.91
			N(-1,1)	.23	.49	.64	.81	.93
			B(3,7)	.30	.60	.74	.89	.96
CAT	Random	100%	N(0,1)	.63	.87	.92	.98	.99
			N(-1,1)	.65	.89	.94	.99	1.00
			B(3,7)	.74	.93	.97	.99	1.00
		50%	N(0,1)	.30	.58	.72	.87	.97
			N(-1,1)	.35	.63	.78	.92	.98
			B(3,7)	.55	.81	.89	.96	.99
		25%	N(0,1)	.08	.28	.42	.69	.87
			N(-1,1)	.12	.34	.51	.73	.92
			B(3,7)	.29	.56	.70	.88	.96
	Strategic	100%	N(0,1)	.47	.74	.84	.94	.99
			N(-1,1)	.34	.59	.72	.86	.96
			B(3,7)	.48	.68	.77	.87	.94
		50%	N(0,1)	.24	.54	.68	.87	.96
			N(-1,1)	.29	.56	.68	.85	.94
			B(3,7)	.44	.68	.78	.90	.97
		25%	N(0,1)	.08	.24	.35	.60	.85
			N(-1,1)	.13	.35	.47	.72	.89
			B(3,7)	.31	.55	.67	.85	.95

Summary and Conclusions

The simulation investigation showed that the MDPP ℓ_z index works quite well in conjunction with normal priors. Importantly, the actual Type I error rates were found to be close to the nominal rates. The power to detect aberrant responses was excellent for the simulated static tests: .99 and higher for 100% random responding and .90 and higher for strategic responding at a Type I error rate of just .01. Even when 50% of the responses were aberrant, power was .84 or higher for random responding and .60 and higher for strategic responding with a .01 Type I error rate. These rates are perhaps the highest ever found in the person fit literature.

The detection rates for the simulated CAT, although not as remarkably high as for the static tests, were nonetheless excellent: .87 and higher for 100% random responding and .59 and higher for 100% strategic responding when the Type I error rate was .05.

In sum, the positive results for controlling the Type I error rate while obtaining high rates of detection of aberrant responding provides a strong justification for applying the new method to actual TAPAS MEPS data. In the next chapter, we describe an investigation of this analysis.

CHAPTER 4: INVESTIGATING THE EFFECTS OF UNMOTIVATED RESPONDING ON VALIDITIES OF TAPAS TESTS WITH U.S. ARMY APPLICANTS

Objective

This chapter describes analyses of actual TAPAS static and CAT responses. Using lessons learned from Chapter 2 and Chapter 3, we aimed at 1) developing approaches to flagging individuals who may have provided random, patterned, or strategic responses, 2) estimating the extent of criterion validity improvement when unmotivated examinees were removed from the database, and 3) determining if flagged individuals had obtained higher TAPAS scores than those responding conscientiously. Specifically, we computed construct and criterion validities of TAPAS scores for a group classified as providing valid scores and a group classified as providing invalid scores. We expected to see lower validities for groups consisting of individuals with (1) many very fast responses, (2) patterned responses, and (3) aberrant response patterns as identified by the ℓ_z appropriateness index. We also compared trait and composite scores for normal and aberrant groups to gauge whether examinees who were suspected of engaging in unmotivated responding were gaining any advantage over those who appeared to respond in accordance with provided instructions.

Approaches to Identifying Unmotivated Respondents

Three approaches to identifying potentially unmotivated responding were taken. First, item response times were utilized because it appears to be impossible to answer an item seriously in less than two seconds. We examined various cut-offs for rapid responding (e.g., less than 1 second, less than 2 seconds, etc.) and various numbers of fast responses (more than 5 fast responses, more than 10 fast responses, etc.) to determine the extent to which fast responding affected observed validities. To better understand the profile of those engaged in rapid responding, we calculated respondents' overall TAPAS testing times and studied its relationship with personality and AFQT scores.

The second approach looked at observed response patterns to identify individuals who were "playing games" and providing patterned answers (e.g., ABABAB or AAAAAA). Currently, TAPAS computes a flag based on the number of times an individual selected the response option A during the 120 item test. The idea behind the flag was that, because the order of two response options is randomized prior to the presentation of each item, too many or too few A responses would be indicative of patterned responding. A limitation of this rather simple approach is that it can only flag AAAAAA or BBBBB patterns, but is insensitive to the alternating response pattern (e.g., ABABAB).

To develop an index that would be sensitive to a wider variety of patterned responding, we first computed the Markov chain transition matrix for each examinee as shown in Figure 1 below. The values in the cells of the Markov matrix indicate the number of times two particular response options were observed on successive trials. For example, if a test having six items had the ABBAAA response pattern, the Markov values in the 2x2 table would be AA=2, AB=1, BA=1, and BB=1. For the 120-item test, an example of AA, BB, BA, and AB counts is shown in Figure 1. Due to randomized ordering of response options in TAPAS, the expected counts each Markov value should be equal to 29.75 or (# items - 1)/4. An overall Markov value can be computed as the sum of (Observed-Expected)^2/Expected values across the four cells. The larger the overall value, the higher the likelihood of patterned responding. As can be seen in the figure, the observed counts for the four Markov cells were not too far from the expected value of 29.75. The overall Markov value for this table is 1.44.

Figure 1. Example of Markov Chain Transition Matrix

.

		Response to Item i+1	
Response to Item i		A	B
	A	26	29
	B	29	35

The third approach utilized the newly developed MDPP ℓ_z appropriateness index. Using the stored item response data, we calculated ℓ_z values for all Army respondents in the criterion database and studied the extent to which using increasingly strict cut off values would influence observed validities of TAPAS scores.

Method

Sample. In our database, 31,996 Army applicants had the necessary data for our empirical analyses. Of those, 15,303 completed a Static version of TAPAS and 16,693 completed the 15D CAT. 71.7% of the final sample were Regular Army, 21.5 % were National Guard, and the remaining 6.8 % were Reserve. Most examinees were male (84.1 %). The racial composition was 64.8% Caucasian, 10.5% African American, 10.3% Hispanic, and 1.6% Asian; 20.7% declined to indicate their race.

Table 11 shows descriptive statistics for the 15 TAPAS dimensions from the Static and CAT versions. In addition to showing the IRT-based trait scores (i.e., raw scores) and their standard deviations, we also present normed means and normed standard deviations for each test version. TAPAS norms were developed using all Army applicants who completed the tests between May 2009 and May 2010; scores for all applicants in the norm samples were scaled to follow the standard normal distribution. As can be seen in Table 11, the normed means and standard deviations for the TAPAS dimensions are still near 0 and 1, respectively, indicating that the examinees in this sample were very similar to the larger groups that were used to norm the tests. Hence it appears that the results should generalize fairly well.

Table 11. Descriptive Statistics for the TAPAS Dimensions in the Army Sample

TAPAS Dimensions	Raw Mean	Raw Standard Deviations	Normed[a] Mean	Normed[a] Standard Deviations
TAPAS Static: Achievement	0.26	0.49	-0.04	0.96
TAPAS Static: Adjustment	0.15	0.58	-0.02	0.98
TAPAS Static: Cooperation	-0.06	0.39	0.03	0.97
TAPAS Static: Dominance	-0.02	0.57	0.02	0.96
TAPAS Static: Even Tempered	0.25	0.48	-0.03	0.97
TAPAS Static: Attention Seeking	-0.25	0.53	0.00	0.98
TAPAS Static: Selflessness	-0.19	0.44	0.00	0.96
TAPAS Static: Intellectual Efficiency	-0.12	0.58	-0.03	0.97
TAPAS Static: Non-Delinquency	0.12	0.45	0.01	0.97
TAPAS Static: Order	-0.37	0.56	0.02	0.97
TAPAS Static: Physical Conditioning	-0.04	0.60	0.04	0.96
TAPAS Static: Self-Control	0.09	0.52	-0.02	0.97
TAPAS Static: Sociability	-0.20	0.58	0.02	0.97
TAPAS Static: Tolerance	-0.27	0.58	-0.04	0.96
TAPAS Static: Optimism	0.28	0.49	0.02	0.96
TAPAS CAT: Achievement	0.17	0.48	0.03	0.98
TAPAS CAT: Adjustment	0.04	0.57	0.07	0.98
TAPAS CAT: Cooperation	-0.05	0.37	0.04	0.98
TAPAS CAT: Dominance	0.03	0.60	-0.01	0.99
TAPAS CAT: Even Tempered	0.18	0.47	0.05	0.97
TAPAS CAT: Attention Seeking	-0.19	0.53	0.02	0.98
TAPAS CAT: Selflessness	-0.23	0.43	-0.07	0.98
TAPAS CAT: Intellectual Efficiency	0.01	0.58	0.06	0.97
TAPAS CAT: Non-Delinquency	0.11	0.46	0.06	0.98
TAPAS CAT: Order	-0.47	0.54	-0.07	0.98
TAPAS CAT: Physical Conditioning	0.05	0.62	0.05	0.97
TAPAS CAT: Self-Control	0.07	0.53	0.01	0.98
TAPAS CAT: Sociability	-0.05	0.60	-0.01	0.99
TAPAS CAT: Tolerance	-0.24	0.57	-0.03	0.98
TAPAS CAT: Optimism	0.15	0.45	0.04	0.97

Note: N(Static) = 15,303; N(CAT) = 16,693; [a] = Raw TAPAS scores were rescaled with respect to test norms developed using Army applicants who took the tests between May 2009 and May 2010.

Criteria. In this research, we focused on four criteria that were important to the Army and were found to correlate with TAPAS scores in past reports. These included the Army Life Questionnaire (ALQ) Attrition Cognitions scale, the ALQ Army Life Adjustment scale, the Army Physical Fitness Test (APFT), and 6-month attrition. Data for the four criteria were collected as part of the Tier One Performance Screen (TOPS; Knapp, Heffner, & White, 2011) research project; the time lag between TAPAS administration and criterion data collection ranged from 6 to 18 months. In addition, we included Armed Forces Qualification Test (AFQT) scores, because the scores are known to correlate with TAPAS Intellectual Efficiency, Order, and Achievement scores. The AFQT is a composite of four subtests from the Armed Services Vocational Aptitude Battery (ASVAB), which all applicants completed at about the same time as TAPAS. Descriptive statistics for the five criteria and correlations with TAPAS scores are shown in Tables 12 and 13.

Table 12. Descriptive Statistics for Criterion Measures and AFQT scores in the Army Sample

TAPAS Version	Criterion	N	Mean	Std. Dev.	Min	Max
Static	Army Life Adjustment	806	4.04	0.66	1	5
	Attrition Cognitions	806	1.55	0.64	1	5
	Army Physical Fitness Test (APFT)	794	247.15	32.65	66	300
	6-month Attrition	3177	0.09	0.29	0	1
	AFQT	15303	56.47	24.15	1	99
CAT	Army Life Adjustment	3963	4.08	0.66	1	5
	Attrition Cognitions	3963	1.52	0.6	1	5
	Army Physical Fitness Test (APFT)	3919	251.47	30.47	120	300
	6-month Attrition	14800	0.09	0.29	0	1
	AFQT	16692	61.51	20.64	3	99

Table 13. Correlations between TAPAS Dimensions, Criteria, and AFQT scores

TAPAS Dimension	Army Life Adjustment	Attrition Cognitions	APFT	6-month Attrition	AFQT
Achievement	.14	-.13	.09	-.01	.09
Adjustment	.09	-.02	.00	-.02	.08
Cooperation	-.01	-.02	-.01	-.01	-.07
Dominance	.14	-.09	.14	-.01	.08
Even Tempered	.04	-.04	-.07	-.01	.06
Attention Seeking	.06	-.04	.07	-.03	.10
Selflessness	-.01	-.04	.00	.03	-.08
Intellectual Efficiency	.11	-.05	.05	-.01	.43
Non-Delinquency	.01	-.03	-.05	.01	.00
Order	.00	.01	.01	.02	-.18
Physical Conditioning	.13	-.05	.27	-.07	.02
Self-Control	.03	-.03	-.02	.00	-.05
Sociability	.02	.00	.03	.00	-.10
Tolerance	.03	-.04	.02	.01	.01
Optimism	.11	-.07	.06	-.02	-.01

Results

TAPAS Scores and Validities for Rapid Responders. First, for each examinee, we recoded TAPAS item response latencies into three mutually exclusive latency bands: “less than 2 seconds," "2 to 8 seconds," and "more than 8 seconds." Then, for each examinee, we counted how many of the response times fell into each of the bands.

Because it is nearly impossible to read and respond carefully to pairwise preference items in less than 2 seconds, applicants with high numbers of responses in the first band are likely to be unmotivated (random) responders. Examinees who take 2 to 8 seconds to respond to a TAPAS may be fast information processors or just decisive in their choices, while those taking more than 8 seconds may read more slowly or prefer to deliberate about their answers. Although examinees with a majority of responses in the latter two latency bands may still engage in other kinds of aberrant responding (e.g., patterned or faking), they were treated as "normal" for the purposes of our rapid response analyses.

Table 14 shows correlations between total testing time, frequency counts for the response latency bands, AFQT scores, and TAPAS scores. As expected, examinees with higher Intellectual Efficiency and AFQT scores had shorter overall testing times and higher counts for responses in the 2-8 second band. However, AFQT correlated negatively with frequency counts for the <2 second band, indicating that lower ability examinees more often engaged in rapid responding. Also, as expected, TAPAS Dominance and Sociability scores negatively correlated with total testing time and frequency counts for the >8 second band, indicating that extraverted individuals tended to answer TAPAS items more decisively. In contrast, Self-Control correlated positively with total testing time and frequency counts for the >8 second band, which is consistent with the notion that individuals high on self-control are careful and tend to think before they act.

Table 14. Correlations between Total Testing Time, Frequencies for the Three Response Latency Bands, AFQT, and TAPAS Scores

Tapas Dimension	Total Testing Time	Response Latency Band		
		<2 seconds	2-8 seconds	>8 seconds
AFQT	-.16	-.08	.21	-.18
Achievement	.03	-.09	.01	.02
Adjustment	.03	-.02	-.03	.04
Cooperation	-.03	.03	.02	-.03
Dominance	-.07	.01	.08	-.08
Even Tempered	.06	-.07	-.04	.06
Attention Seeking	-.05	.04	.03	-.04
Selflessness	.04	.02	-.03	.02
Intellectual Efficiency	-.11	-.03	.14	-.13
Non-Delinquency	.02	-.02	.00	.01
Order	.01	.07	-.03	.01
Physical Conditioning	-.08	.03	.07	-.08
Self-Control	.11	-.03	-.09	.10
Sociability	-.10	.03	.09	-.10
Tolerance	.03	-.01	-.01	.02
Optimism	.00	-.03	.01	.00

Table 15 shows the frequency counts for the number of items each examinee answered in less than two seconds. For example, the frequency count for “0” items was 28,443, indicating that that 88.9% of the examinees finished the test without a single instance of rapid responding. In fact, only 5% of examinees in the sample had four or more instances of rapid responding. Hence, the occurrence of this type of aberrant responding appears to be fairly low.

Table 15. Frequency Distribution for “Less than 2 Seconds” Response Latencies

# Items Answered in < 2 Seconds	Frequency	Percent	Cumulative Percent
0	28,443	88.9	88.9
1	1,300	4.06	93.0
2	416	1.30	94.3
3	248	0.78	95.0
4	168	0.53	95.6
5	123	0.38	95.9
6	65	0.20	96.1
7	82	0.26	96.4
8	68	0.21	96.6
9	58	0.18	96.8
10	48	0.15	96.9
11 to 20	277	0.87	97.8
21 to 30	177	0.55	98.4
31 to 40	97	0.30	98.7
41 to 50	81	0.25	98.9
51 to 60	82	0.26	99.2
61 to 70	59	0.18	99.4
71 to 80	42	0.13	99.5
81 to 90	34	0.11	99.6
91 to 100	49	0.15	99.8
101 to 110	36	0.11	99.9
111 to 120	43	0.13	100.0

Given the frequency distribution for “less than 2 seconds” responding, shown in Table 15, and the results of the simulation studies, described in Chapters 2 and 3, we classified examinees who answered more than 12 items with less than 2 seconds response latencies as rapid responders. Table 16 shows descriptive statistics for the 15 TAPAS dimensions, the TAPAS Can Do and Will Do selection composites, criterion measures, and AFQT scores. In the table, we compare three groups of examinees: those answering 0, 1 to 12, and more than 12 items in less than 2 seconds each. The first two groups were considered “normal” responders and the last group was designated as rapid responders. As can be seen, scores for the rapid responders were lower for many TAPAS dimensions and the Can Do and Will Do composites. Thus, they are more likely to fail the test than those who responded more diligently. Note that the occasional occurrence of rapid responding, which is signified by the “1-12 Items” group, had generally minor effects on TAPAS means, so the Can Do and Will Do composite means were very similar to the “0 Items” group. The rapid responders also had lower AFQT and Army Life Adjustment scores and slightly higher 6-month attrition.

Table 16. Means and Standard Deviations for Normal and Rapid Responder Groups

Dimension	Applicants with "Less than 2 Seconds" Response Times					
	0 Items		1-12 Items		> 12 Items	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Achievement	0.01	0.98	-0.02	0.90	-0.45	0.73
Adjustment	0.04	0.99	0.00	0.95	-0.03	0.84
Cooperation	0.01	0.98	0.21	0.95	0.24	0.88
Dominance	-0.02	0.99	0.21	0.93	0.08	0.64
Even Tempered	0.03	0.98	-0.03	0.93	-0.34	0.77
Attention Seeking	-0.01	0.98	0.14	1.00	0.24	0.77
Selflessness	-0.03	0.98	-0.10	0.89	0.02	0.78
Intellectual Efficiency	0.02	0.98	0.01	0.91	-0.12	0.72
Non-Delinquency	0.03	0.97	0.14	1.00	-0.08	0.84
Order	-0.06	0.98	0.19	0.90	0.36	0.63
Physical Conditioning	0.02	0.98	0.27	0.90	0.22	0.61
Self-Control	0.00	0.98	-0.06	0.96	-0.19	0.79
Sociability	-0.03	0.98	0.28	0.93	0.27	0.75
Tolerance	-0.03	0.99	-0.11	0.86	-0.11	0.74
Optimism	0.02	0.98	0.13	0.89	-0.06	0.78
Will Do Composite	0.04	1.00	0.09	0.99	-0.38	0.85
Can Do Composite	0.04	1.00	0.08	0.96	-0.40	0.84
Army Life Adjustment	4.08	0.65	4.02	0.76	3.86	0.74
Attrition Cognitions	1.52	0.60	1.60	0.69	1.61	0.63
APFT	250.87	30.90	250.28	31.44	245.57	26.95
6-month Attrition	0.09	0.29	0.09	0.29	0.11	0.31
AFQT	60.23	22.25	51.61	22.70	45.65	21.97

Note: N(0 Items) = 28,443; N(1-12 Items) = 2,617; N(>12 Items)=936. For the criteria (6-month Attrition, APFT, Attrition Cognitions, and Army Life Adjustment) and AFQT, sample sizes ranged from 28,443 down to 69.

Table 17 presents correlations between TAPAS scores and AFQT and the criteria for the three comparison groups: 0, 1-12, and >12 items answered in less than 2 seconds. For rapid responders (> 12 items), the correlation of Physical Conditioning with APFT was .00, whereas it was .29 in the 0 Items group. In addition, Intellectual Efficiency correlated only .17 with AFQT among rapid responders versus .44 in the 0 Items group. Overall, many correlations among rapid responders were inconsistent with theoretical expectations, suggesting that response time screens may be useful when examining TAPAS validities, especially for criteria with small samples.

Table 17. Criterion Correlations for TAPAS Scores Among Normal and Rapid Responder Groups

	Army Life Adjustment			Attrition Cognitions			APFT			6-month Attrition			AFQT		
	0 Items	1-12 Items	>12 Items	0 Items	1-12 Items	>12 Items	0 Items	1-12 Items	>12 Items	0 Items	1-12 Items	>12 Items	0 Items	1-12 Items	>12 Items
Sample Size	4,337	363	69	4,337	363	69	4,284	360	69	16,416	1,268	293	28,442	2,617	936
Achievement	.15	.12	-.08	-.14	-.11	.08	.09	.09	.08	-.01	.00	.02	.08	.08	.00
Adjustment	.10	.04	-.04	-.03	.02	.24	.01	-.05	-.15	-.02	-.02	.08	.09	.04	.05
Cooperation	-.02	.05	.02	-.01	-.08	-.07	.00	-.06	-.15	-.01	.06	-.03	-.06	-.10	.03
Dominance	.15	.09	.00	-.10	-.02	-.03	.15	.02	.18	-.01	-.06	-.04	.09	.06	.06
Even Tempered	.04	.07	.02	-.04	-.06	.07	-.07	-.13	.04	-.02	.00	.17	.05	.05	.08
Attention Seeking	.06	.02	-.13	-.04	-.07	.07	.08	-.09	.03	-.02	-.04	.00	.11	.10	.11
Selflessness	-.01	.01	.02	-.04	-.11	-.05	-.01	.02	.09	.03	.03	.02	-.08	-.05	.05
Intellectual Efficiency	.12	.05	.02	-.05	-.06	-.06	.05	-.03	.05	-.02	.02	-.09	.44	.42	.17
Non-Delinquency	.01	.00	-.19	-.03	-.04	-.03	-.05	-.09	-.06	.01	.05	.06	.00	.00	.03
Order	.00	.05	-.07	.01	.06	-.05	.01	.10	.01	.02	-.02	.02	-.16	-.18	-.13
Physical Conditioning	.14	.03	-.04	-.06	-.02	-.14	.29	.14	.00	-.07	-.05	-.04	.03	.01	.02
Self-Control	.02	.11	.19	-.02	-.06	-.14	-.02	.08	.05	.00	.00	.10	-.05	-.09	-.03
Sociability	.03	-.05	-.18	-.01	.07	.04	.04	-.02	.08	-.01	.02	.08	-.09	-.06	-.02
Tolerance	.03	.01	.09	-.04	-.04	-.08	.02	-.01	.08	.00	.08	.12	.00	.02	.03
Optimism	.13	-.06	-.06	-.07	-.06	.05	.07	.01	-.07	-.03	.00	-.04	-.01	-.01	.09
Will Do	.11	.07	-.06	-.09	-.06	-.02	.07	.04	.01	-.03	.02	.08	.02	.01	.01
Can Do	.16	.06	-.10	-.12	-.12	.04	.03	-.06	.01	-.02	.03	.05	.21	.19	.13

TAPAS Scores and Validities for Patterned Responders. The first step in attempting to identify patterned responders was to determine critical values for the Markov index having designated false positive rates; i.e., if an examinee in the Army sample had an observed Markov value greater than that critical value, he or she would be classified as a patterned responder. To determine these critical values, we simulated data for 1,000 normal responders, computed their Markov values, ranked the values in ascending order, and identified those corresponding to the 90th to 99th percentiles. These results are shown in Table 18 under the column heading Simulated Normal Responders.

As can be seen in Table 18, a critical value of 16.56 would result in 1% of normal examinees being misclassified as patterned responders and a value of 9.43 would lead to a 5% false positive rate. After obtaining the critical values, we calculated Markov values for the 31,996 examinees in the Army sample and found that 925 examinees had Markov values greater than 16.56 and another 1,899 examinees had Markov values between 9.43 and 16.56. Using these cutoffs, we divided the Army sample into three mutually exclusive groups for comparisons of TAPAS score means, standard deviations, and criterion correlations.

Table 18. Descriptive Statistics and Selected Percentiles for Markov Values in the Simulated Normal Sample and the Army Sample

Percentiles	Simulated Normal Responders	Army Sample
N	1,000	31,996
Mean	3.09	4.58
Std. Dev.	3.45	14.36
Minimum	0.03	0.03
Maximum	31.06	357
90	7.14	8.83
91	7.68	9.3
92	8.15	9.91
93	8.49	10.58
94	8.97	11.52
95	9.43	12.66
96	10.3	14.14
97	12.33	16.03
98	13.87	19.18
99	16.56	27.59

Table 19 presents descriptive statistics for the 15 TAPAS dimensions, the TAPAS Can Do and Will Do selection composites, AFQT and the criteria for the three comparison groups defined by the observed Markov values. The first group (<9.43) was considered normal for comparison purposes, while the other groups exhibited greater degrees of patterned responding. As can be seen in the table, the group with Markov values >16.56 had lower means for Achievement, Even Tempered, and Intellectual Efficiency dimensions, which translated into lower overall means on the Can Do and Will Do composites. These findings are similar to what was observed for rapid responders, although the overall effects are smaller. Note that the TAPAS means and standard deviations for the second group (9.43 to 16.56) are very similar to those for the <9.43 group, suggesting that 9.43 critical value might be flagging too many normal examinees. Also note that the criterion means were fairly similar across groups, with the exception of AFQT, which showed a small but clear trend for higher scores among persons with lower Markov values.

Table 19. Means and Standard Deviations for Normal and Patterned Responder Groups

	Markov Value in Army Sample					
	<9.43		9.43 - 16.56		>16.56	
Dimension	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Achievement	0.01	0.98	-0.07	0.92	-0.25	0.87
Adjustment	0.04	0.98	-0.01	0.98	-0.09	0.93
Cooperation	0.03	0.98	0.11	0.96	0.18	0.91
Dominance	0.00	0.99	0.01	0.90	0.05	0.79
Even Tempered	0.02	0.98	0.01	0.96	-0.17	0.89
Attention Seeking	0.01	0.99	0.01	0.96	0.06	0.81
Selflessness	-0.04	0.98	0.00	0.93	0.04	0.85
Intellectual Efficiency	0.02	0.97	-0.06	0.99	-0.10	0.87
Non-Delinquency	0.04	0.98	0.00	0.95	-0.03	0.91
Order	-0.04	0.98	0.03	0.93	0.19	0.84
Physical Conditioning	0.04	0.97	0.03	0.92	0.11	0.76
Self-Control	0.00	0.98	-0.03	0.96	-0.09	0.87
Sociability	0.00	0.99	0.02	0.92	0.10	0.85
Tolerance	-0.04	0.98	-0.04	0.91	-0.04	0.84
Optimism	0.03	0.97	0.02	0.94	-0.04	0.84
Will Do Composite	0.04	1.00	-0.02	0.99	-0.17	0.93
Can Do Composite	0.05	0.99	-0.04	0.97	-0.22	0.94
Army Life Adjustment	4.07	0.66	4.03	0.71	4.06	0.58
Attrition Cognitions	1.53	0.61	1.56	0.64	1.48	0.51
APFT	250.80	30.88	249.00	31.38	252.84	29.97
6-month Attrition	0.09	0.29	0.09	0.29	0.09	0.29
AFQT	59.55	22.36	55.49	23.58	52.17	23.81

Note: N (<9.43) = 29,172; N (9.43-16.56) = 1,899; N(>16.56) =925.

Table 20 presents correlations between TAPAS scores, AFQT, and the criteria for the three Markov groups: <9.43, 9.43-16.56, >16.56. Unlike the results for rapid responders which showed some deterioration in validities for the high group (more than 12 items answered in less than two seconds each) relative to the “normal” examinees, there was little evidence here that the high Markov group (>16.56) had lower validities than the other groups. The correlations of Intellectual Efficiency with AFQT and Physical Conditioning with APFT were fairly similar across the three groups, which suggests that the critical value of 16.56 still may have flagged too many normal examinees as patterned responders. To explore this possibility further, we conducted an additional analysis on the Army sample using a stringent critical value of 31.06, which was the maximum observed in the simulated data ($p < .001$). Out of the 31,996 Army examinees, 264 had Markov values exceeding this threshold. Table 21 presents the TAPAS means, standard deviations, and correlations with AFQT for this extreme group. For convenience, we also included the results for the low Markov (<9.43) group.

As can be seen in Table 21, the means and standard deviations for the >31.06 group were considerably lower than the <9.43 group and, importantly, the Can Do and Will Do composite means were nearly one standard deviation lower. In addition, the correlation between Intellectual Efficiency and AFQT dropped from .43 to .07. Thus, with the pronounced differences across these comparison groups, we recommend using a high Markov value, such as 31.06, for data screening purposes.

Table 20. Criterion Correlations for TAPAS Scores Among Normal and Patterned Responder Groups

	Army Life Adjustment			Attrition Cognitions			APFT			6-month Attrition			AFQT		
	<9.43	9.43-16.56	>16.56	<9.43	9.43-16.56	>16.56	<9.43	9.43-16.56	>16.56	<9.43	9.43-16.56	>16.56	<9.43	9.43-16.56	>16.56
Sample Size	4,367	285	117	4,367	285	117	4,312	284	117	16,370	1,120	487	2,9171	1,899	925
Achievement	.13	.21	.15	-.13	-.22	-.02	.09	.12	.08	-.01	.00	.02	.08	.08	.09
Adjustment	.10	.01	.03	-.03	.07	-.03	.01	.02	-.09	-.02	.03	-.05	.08	.16	.10
Cooperation	-.01	-.02	.03	-.01	-.01	-.07	-.01	.05	-.05	.00	-.02	.00	-.07	-.08	-.05
Dominance	.14	.14	.21	-.09	-.14	-.10	.14	.03	.16	-.01	-.01	-.08	.08	.04	.04
Even Tempered	.04	.03	.04	-.04	-.04	-.03	-.08	.04	-.11	-.01	-.01	-.02	.06	.09	.12
Attention Seeking	.06	.00	.20	-.04	-.01	-.10	.07	.04	.13	-.02	-.05	-.07	.10	.10	.05
Selflessness	.00	-.25	.19	-.05	.12	-.10	-.01	.01	.03	.03	.05	.08	-.07	-.12	-.10
Intellectual Efficiency	.11	.09	.10	-.05	-.05	-.05	.05	.07	-.08	-.01	.01	-.06	.43	.46	.37
Non-Delinquency	.01	-.04	.00	-.03	.04	-.07	-.05	-.03	-.16	.01	.04	.07	.00	.00	-.02
Order	.00	-.07	.10	.01	.13	-.10	.02	-.02	-.01	.02	.05	.02	-.17	-.19	-.21
Physical Conditioning	.13	.07	-.05	-.05	-.05	.06	.28	.24	.22	-.07	-.04	-.11	.02	-.02	.01
Self-Control	.03	.00	.07	-.03	.03	-.03	-.01	-.05	.00	.00	.10	.06	-.04	-.13	-.07
Sociability	.02	-.01	.09	.00	.01	.09	.03	.06	-.01	.00	-.03	-.03	-.10	-.06	-.06
Tolerance	.03	.01	.13	-.04	-.07	.00	.02	.07	.08	.01	.05	.03	.01	.02	.09
Optimism	.12	.04	.01	-.07	-.02	-.02	.06	.09	.06	-.02	-.05	.03	-.01	-.03	.03
Will Do	.11	.11	-.01	-.09	-.10	.01	.07	.12	-.05	-.02	.02	.01	.02	.02	.06
Can Do	.15	.12	.10	-.12	-.10	-.07	.02	.10	-.08	-.02	.00	.02	.20	.22	.20

Table 21. Means, Standard Deviations, and Correlations for High (>31.06) and Low (<9.43) Markov Groups

TAPAS Dimension	Markov Value > 31.06			Markov Value < 9.43		
	Mean	Std. Dev.	AFQT Correlation	Mean	Std. Dev.	AFQT Correlation
Achievement	-.64	.69	-.13	.01	.98	.08
Adjustment	-.21	.82	-.19	.04	.98	.08
Cooperation	.24	.84	.02	.03	.98	-.07
Dominance	.08	.58	.11	.00	.99	.08
Even Tempered	-.46	.67	-.07	.02	.98	.06
Attention Seeking	.21	.63	.05	.01	.99	.10
Selflessness	.09	.70	.06	-.04	.98	-.07
Intellectual Efficiency	-.31	.72	.07	.02	.97	.43
Non-Delinquency	-.16	.71	.03	.04	.98	.00
Order	.33	.56	-.02	-.04	.98	-.17
Physical Conditioning	.12	.55	.03	.04	.97	.02
Self-Control	-.10	.79	-.09	.00	.98	-.04
Sociability	.06	.62	-.11	.00	.99	-.10
Tolerance	-.12	.64	.02	-.04	.98	.01
Optimism	-.20	.73	-.03	.03	.97	-.01
Will Do	-.57	.75	-.08	.04	1.00	.02
Can Do	-.66	.74	-.05	.05	.99	.20

Note: N (<9.43) = 29,172; N(>31.06) =264.

TAPAS Scores and Validities for Responders with Low ℓ_z Values. Figure 2 shows the distributions of ℓ_z values for Army examinees who took the Static and CAT versions of TAPAS. Both ℓ_z distributions are approximately normal, but the means and standard deviations differed somewhat from the theoretical values of zero and 1, respectively, with CAT having a higher mean and a slightly smaller standard deviation.

Figure 2. Distribution of ℓ_z Values for Static and CAT TAPAS Versions.

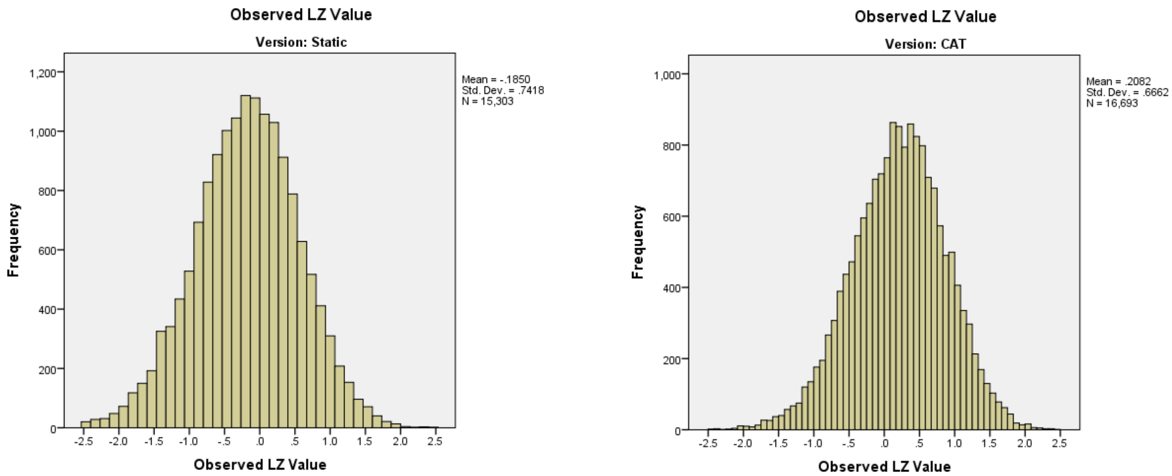

Table 22 shows descriptive statistics for ℓ_z values among Army examinees who took the Static and CAT versions of TAPAS. First, note that the minimum and maximum ℓ_z values differed markedly across the tests, with Static showing a much wider range than CAT. The reduced range for CAT might be explained by the statistical arguments concerning power to detect inconsistencies with model predictions as adaptive testing progresses (see Chapter 3). It is also possible that the Static test showed a wider range of ℓ_z values because it was introduced first into the MEPS and some examinees may have taken the test less seriously in the early stages of the research trial.

Table 22. Descriptive Statistics and Percentiles for ℓ_z Values in the Army Sample

Percentiles	Static	CAT
N	15,303	16,693
Mean	-.19	.21
Std. Dev.	.74	.67
Minimum	-3.32	-2.74
Maximum	4.87	2.47
1	-2.00	-1.47
2	-1.77	-1.23
3	-1.63	-1.09
4	-1.51	-0.99
5	-1.43	-0.91
6	-1.37	-0.84
7	-1.30	-0.78
8	-1.25	-0.73
9	-1.19	-0.69
10	-1.14	-0.65

In the Monte Carlo simulation described in Chapter 3, we proposed and evaluated the efficacy of person-specific ℓ_z critical values for classifying examinees as normal or aberrant. The results showed good to excellent power to detect aberrant responding in conjunction with Type I errors very close to the nominal alpha levels. However, when applying that methodology to the response data for the Static and CAT versions of TAPAS taken by Army examinees, exceedingly high proportions of examinees were flagged as aberrant: 41.7% for Static and 21.6% of CAT. This suggested that the simulation-based person-specific ℓ_z critical values might not be robust to the violations of model assumptions that are likely with real data. Consequently, we examined the observed ℓ_z distributions for Static and CAT versions of TAPAS, shown in Table 22, and chose two group-level ℓ_z critical values for each test to use for classification.

As in the previous analyses involving random and patterned responders, we created three examinee groups for comparison. The first group (>5%) had ℓ_z values above the 5th percentile based on the version of TAPAS they took (i.e., $\ell_z > -1.43$ for Static and $\ell_z > -.91$ for CAT). The second group had ℓ_z values between the 2nd and 5th percentiles (-1.77 to -1.43 for Static and -1.23 to -.91 for CAT), and the third group had ℓ_z values below the 2nd percentile (< -1.77 for Static and < -1.23 for CAT).

Table 23 presents descriptive statistics for the 15 TAPAS dimensions, the TAPAS Can Do and Will Do selection composites, AFQT, and the four criterion measures for the three comparison groups defined by the observed ℓ_z values. The first group (>5%) was considered normal for comparison purposes, while the other groups exhibited increasing degrees of aberrant responding. As can be seen in the table, the 2% - 5% and <2% groups had lower means and standard deviations on Achievement, Even Tempered, Non-Delinquency, and on the Will Do and Can Do composites. However, the overall declines in composite scores were smaller than those observed in the analyses for rapid and patterned responding, perhaps because ℓ_z is sensitive to both random and strategic responding, which have different effects on TAPAS scores (see the simulation results in Chapters 2 and 3).

*Table 23. Means and Standard Deviations for **ℓ_z** Groups*

	Applicants ℓ_z Value					
	> 5%		2% - 5%		<2%	
Dimension	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Achievement	0.00	0.98	-0.12	0.79	-0.19	0.69
Adjustment	0.03	0.99	0.05	0.80	0.09	0.78
Cooperation	0.04	0.99	0.01	0.79	-0.04	0.78
Dominance	0.00	1.00	0.02	0.64	0.01	0.60
Even Tempered	0.02	0.98	-0.17	0.81	-0.19	0.70
Attention Seeking	0.00	0.99	0.18	0.75	0.17	0.67
Selflessness	-0.04	0.98	0.04	0.79	0.06	0.69
Intellectual Efficiency	0.02	0.98	-0.09	0.69	-0.08	0.59
Non-Delinquency	0.05	0.98	-0.13	0.79	-0.12	0.71
Order	-0.04	0.98	0.25	0.80	0.28	0.69
Physical Conditioning	0.04	0.98	0.04	0.66	0.02	0.61
Self-Control	0.00	0.99	-0.08	0.79	-0.13	0.72
Sociability	0.00	0.99	0.09	0.72	0.09	0.63
Tolerance	-0.04	0.98	0.00	0.77	0.05	0.67
Optimism	0.04	0.97	-0.06	0.75	-0.14	0.69
Will Do Composite	0.05	1.00	-0.23	0.83	-0.27	0.75
Can Do Composite	0.05	1.00	-0.21	0.81	-0.27	0.71
Army Life Adjustment	4.07	0.66	4.04	0.68	4.09	0.68
Attrition Cognitions	1.53	0.61	1.53	0.63	1.59	0.61
APFT	250.95	30.98	248.94	26.36	243.76	31.67
6-month Attrition	0.09	0.29	0.09	0.29	0.10	0.31
AFQT	59.57	22.47	51.06	21.78	48.64	21.35

Note: N (>5%) = 30397; N (2%-5%) = 960; N(<2%) =639.

Table 24 presents correlations between TAPAS scores, AFQT, and the criteria for the three ℓ_z groups: >5%, 2% - 5%, <2%. Overall, there were no consistent differences in validities across the groups. Although the correlations of Intellectual Efficiency with AFQT and Physical Conditioning with APFT declined for the <2% group relative to the >5% group, the correlations between TAPAS Adjustment and Army Life Adjustment actually increased slightly and validities for the other criteria were generally similar. At this point, it is unclear why validities among examinees with the higher ℓ_z values did not decline as expected. It could be the case that sampling error, due to the small Ns for some criteria, obscured the true relationships. Alternatively, because the ℓ_z index can flag both random and strategic responders, it is likely that the 2% - 5% and <2% groups contained a mix of examinees with differing motivations. On one hand, examinees with a poor attitude toward testing might have responded randomly to TAPAS items and received lower scores. And if that same attitude carried into basic training, they would likely have received lower criterion scores. On the other hand, highly motivated examinees may have tried to respond strategically to TAPAS items in an effort to raise their scores, and if that motivation continued through basic training, they may have earned higher criterion scores. Thus, the correlations between TAPAS scores and criterion scores could have actually increased.

Because the results in Tables 23 and 24 were not immediately helpful in identifying subgroups of examinees among which TAPAS validities showed the initially expected patterns of decline, we recommend using a very conservative cutoff for ℓ_z response screening. For the analyses in the next section, we therefore used ℓ_z critical values for Static and CAT versions of TAPAS corresponding to the 2^{nd} percentile.

Table 24. Criterion Correlations for TAPAS Scores Among Groups with Different Observed ℓ_z Values

	Army Life Adjustment			Attrition Cognitions			APFT			6-month Attrition			AFQT		
	>5%	2%-5%	<2%	>5%	2%-5%	<2%	>5%	2%-5%	<2%	>5%	2%-5%	<2%	>5%	2%-5%	<2%
Sample Size	4,538	136	95	4,538	136	95	4,484	134	95	17,088	544	345	30,396	960	639
Achievement	.14	.11	.22	-.13	-.17	-.01	.09	.06	-.06	-.01	-.02	-.01	.09	.07	.10
Adjustment	.09	.17	.18	-.02	.02	-.12	.01	-.05	.00	-.02	-.04	.00	.09	.01	.04
Cooperation	-.02	.10	.32	-.02	.04	-.09	-.01	-.02	.00	-.01	.00	-.02	-.07	-.12	.00
Dominance	.14	.11	.30	-.09	-.05	-.15	.14	.05	-.07	-.01	-.05	.04	.08	.01	.04
Even Tempered	.04	.06	.14	-.03	-.09	-.11	-.07	-.21	-.25	-.01	-.04	.01	.06	.09	.10
Attention Seeking	.06	-.05	-.23	-.04	-.04	.24	.07	.16	.11	-.03	.03	-.09	.11	.11	.07
Selflessness	-.01	.02	.10	-.04	-.10	-.31	.00	-.06	-.16	.03	-.04	.04	-.08	-.04	.02
Intellectual Efficiency	.11	.10	.17	-.05	.05	-.05	.05	-.08	-.01	-.01	.00	.02	.44	.28	.28
Non-Delinquency	.00	.08	.11	-.03	-.07	-.15	-.05	-.11	.00	.01	.02	.04	.00	.00	.01
Order	.00	.01	-.23	.01	.01	.10	.02	.00	.01	.02	.03	.01	-.17	-.20	-.15
Physical Conditioning	.12	.18	.23	-.05	.03	-.06	.28	.24	.07	-.07	.02	-.06	.02	.00	-.01
Self-Control	.03	.04	.17	-.03	-.03	-.12	-.01	-.08	-.09	.00	.03	.03	-.05	-.11	-.01
Sociability	.01	.15	.25	.00	-.05	-.06	.03	.10	-.08	.00	-.04	.12	-.10	-.10	-.02
Tolerance	.03	.08	.04	-.04	-.04	-.09	.02	.03	.04	.01	-.07	.06	.01	.05	.01
Optimism	.11	.03	.02	-.07	-.15	-.10	.06	.03	-.05	-.03	-.03	.08	-.02	.02	.01
Will Do	.10	.17	.32	-.08	-.10	-.20	.07	-.08	-.13	-.02	-.02	.03	.02	.03	.05
Can Do	.15	.14	.21	-.11	-.16	-.14	.03	-.11	-.13	-.02	-.02	.05	.20	.16	.17

Studying the Overall Effects of Aberrant Responding on TAPAS Scores and Validities. In the previous sections of this chapter, we identified three flags for classifying response patterns as aberrant. First, we suggested flagging examinees who answered more than 12 items with less than 2 second response latencies. Second, flag examinees with Markov values greater than 31.06 in an effort to detect pattern responding. Third, flag examinees with ℓ_z values less than -1.77 on Static TAPAS and less than -1.23 on CAT TAPAS to detect examinees who responded inconsistently with model predictions.

In the analyses that follow, we used these flags in conjunction with each other to classify each examinee in the Army sample as either a normal or aberrant responder and then we examined how removing the aberrant cases affected TAPAS scores and validities. Specifically, an examinee was designated as an aberrant responder if he or she was flagged by any of the three aberrance thresholds.

Table 25 shows the frequency counts for normal and aberrant responders based on the simultaneous use of the three aberrance flags. As can be seen in the table, 30,470 examinees were designated as normal and the remaining 1,526 examinees were flagged as providing aberrant responses by at least one of the indices. Only 24 examinees were identified as aberrant by all three.

Table 25. Frequency Counts across Three Types of Aberrance Flags

Markov Value	# Items in <2 Seconds	ℓ_z Value	
		Normal	<2%
Normal	Normal	30470	539
	>12 Items	649	74
>30.06	Normal	49	2
	>12 Items	189	24

Table 26 presents descriptive statistics for the 15 TAPAS dimensions, the TAPAS Can Do and Will Do selection composites, AFQT, and the four criterion measures for the total Army sample (N = 31,996), the aberrant sample (N = 1,526), and a "clean" sample (N = 30,470), which excluded the responses that were designated as aberrant. We also show effect size statistics indicating the differences between the clean and total samples. Because the number of aberrant cases was low relative to the size of the total sample, removing them had little effect on the overall TAPAS score means; none of the TAPAS scores or composites changed by more than .02 standard deviations. Similarly, as shown in Table 27, criterion validities for the clean sample were nearly identical to those in the total sample. These findings echoed simulation results from Chapter 2, which showed that even greater proportions of aberrant responding had little effect on TAPAS scores and validities.

Table 26. Comparisons of Means and Standard Deviations for Total, Aberrant, and Clean Samples

Dimension	Total Sample		Aberrant Sample		Clean Sample		Effect Size
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	
Achievement	0.00	0.97	-0.33	0.74	0.01	0.98	0.02
Adjustment	0.03	0.98	0.02	0.84	0.03	0.99	0.00
Cooperation	0.04	0.98	0.14	0.86	0.03	0.98	-0.01
Dominance	0.00	0.98	0.07	0.64	0.00	0.99	0.00
Even Tempered	0.01	0.97	-0.27	0.76	0.02	0.98	0.01
Attention Seeking	0.01	0.98	0.21	0.75	0.00	0.99	-0.01
Selflessness	-0.03	0.97	0.04	0.76	-0.04	0.98	0.00
Intellectual Efficiency	0.02	0.97	-0.11	0.69	0.02	0.98	0.01
Non-Delinquency	0.04	0.97	-0.09	0.80	0.04	0.98	0.01
Order	-0.03	0.97	0.32	0.67	-0.05	0.98	-0.02
Physical Conditioning	0.04	0.97	0.14	0.64	0.04	0.98	-0.01
Self-Control	-0.01	0.98	-0.14	0.78	0.00	0.98	0.01
Sociability	0.01	0.98	0.20	0.72	0.00	0.99	-0.01
Tolerance	-0.04	0.97	-0.04	0.72	-0.04	0.98	0.00
Optimism	0.03	0.96	-0.07	0.76	0.04	0.97	0.01
Will Do Composite	0.03	0.99	-0.32	0.83	0.05	1.00	0.02
Can Do Composite	0.03	0.99	-0.33	0.81	0.05	1.00	0.02
Army Life Adjustment	4.07	0.66	3.99	0.71	4.07	0.66	0.00
Attrition Cognitions	1.53	0.61	1.61	0.63	1.53	0.61	0.00
APFT	250.74	30.89	245.25	29.59	250.94	30.91	0.01
6-month Attrition	0.09	0.29	0.11	0.31	0.09	0.29	0.00
AFQT	59.10	22.53	47.17	21.93	59.70	22.39	0.03

Note: N (Total) = 31,996; N (Aberrant) = 1,526; N(Clean) =30,470; Effect Size = ($Mean_{Clean}$-$Mean_{Total}$)/$Std.Dev._{Clean}$.

Table 27. Comparison of Criterion Correlations for TAPAS Scores across Total, Clean, and Aberrant Samples

TAPAS Dimension	Army Life Adjustment			Attrition Cognitions			APFT Score			6-month Attrition			AFQT		
	Total	Clean	Diff.	Total	Clean	Diff.	Total	Clean	Diff.	Total	Clean	Diff.	Total	Clean	Diff.
Sample Size	4,769	4,608		4,769	4,608		4,713	4,552		17,977	17,343		31,995	30,469	
Achievement	.14	.14	.00	-.13	-.14	.00	.09	.09	.00	-.01	-.01	.00	.09	.08	-.01
Adjustment	.09	.09	.00	-.02	-.03	.00	.00	.01	.00	-.02	-.02	.00	.08	.08	.00
Cooperation	-.01	-.02	-.01	-.02	-.01	.00	-.01	-.01	.00	-.01	.00	.00	-.07	-.07	.00
Dominance	.14	.14	.00	-.09	-.09	.00	.14	.14	.00	-.01	-.01	.00	.08	.08	.00
Even Tempered	.04	.04	.00	-.04	-.04	.00	-.07	-.07	.00	-.01	-.01	.00	.06	.05	-.01
Attention Seeking	.06	.06	.01	-.04	-.04	-.01	.07	.07	.00	-.03	-.03	.00	.10	.11	.01
Selflessness	-.01	-.01	.00	-.04	-.04	.00	.00	.00	.00	.03	.03	.00	-.08	-.08	.00
Intellectual Efficiency	.11	.11	.00	-.05	-.05	.00	.05	.05	.00	-.01	-.01	.00	.43	.44	.01
Non-Delinquency	.01	.01	.00	-.03	-.03	.00	-.05	-.05	.00	.01	.01	.00	.00	.00	.00
Order	.00	.01	.01	.01	.01	.00	.01	.02	.00	.02	.02	.00	-.18	-.17	.01
Physical Conditioning	.13	.13	.00	-.05	-.05	.00	.27	.28	.01	-.07	-.07	.00	.02	.02	.00
Self-Control	.03	.03	.00	-.03	-.03	.00	-.02	-.01	.00	.00	.00	.00	-.05	-.05	-.01
Sociability	.02	.02	.00	.00	.00	.00	.03	.04	.00	.00	-.01	.00	-.10	-.09	.00
Tolerance	.03	.03	.00	-.04	-.04	.00	.02	.02	.00	.01	.01	.00	.01	.01	.00
Optimism	.11	.11	.00	-.07	-.07	.00	.06	.06	.00	-.02	-.03	.00	-.01	-.02	-.01
Will Do	.11	.10	.00	-.09	-.08	.00	.07	.07	.00	-.02	-.02	.00	.03	.02	-.01
Can Do	.15	.15	.00	-.11	-.12	.00	.03	.03	.00	-.02	-.02	.00	.21	.20	-.01

CHAPTER 5: DETECTING 100% RANDOM RESPONDING ON OPERATIONAL TAPAS TESTS

Objective and Design

One important question stemming from the simulation and empirical studies described in Chapters 3 and 4 was how useful the ℓ_z statistic might be for detecting aberrant responding on operational TAPAS tests, with the primary concern being random responding on all or nearly all TAPAS items. Identification of such response patterns accompanied by actions that might lead, for example, to retesting or temporary disqualification could serve as an effective deterrent to unmotivated responding by future examinees.

To examine how effective ℓ_z is in detecting the most extreme form of random responding, we simulated 100% random response patterns for static and CAT versions of TAPAS (N=1000 each) having exactly the same design specifications as the most recent MEPS tests. These simulated random response patterns were mixed with operational test data (N=15,303 Static and N=16,693 CAT) and ℓ_z statistics were computed for the examinees in the respective combined samples. We then calculated a receiver operating curve (ROC) for each combined sample to see whether ℓ_z could differentiate the simulated 100% random responders from the real test takers. Because it is unlikely that an appreciable number of actual respondents engage in completely random responding, we expected to see "fat" ROC curves, which rise sharply above the diagonal signifying equal proportions of hits and false positives.

Results

Figure 3 shows the ℓ_z -based ROC curve for the 120-item static TAPAS test. ℓ_z provided very high sensitivity and specificity as indicated by the nearly right-angle shape of the ROC. Specifically, 99% of random responders had ℓ_z values below the cutoff score of -2.27, but none of the actual examinees were flagged. Thus, ℓ_z appears to be very effective for identifying completely random response patterns associated with Static TAPAS forms.

Figure 4 shows the ℓ_z -based ROC curve for the 120-item CAT TAPAS. As was found in the Chapter 3 simulations, ℓ_z was still effective for detecting 100% random responding with CAT, but the ratio of hits to false positives was smaller than for Static TAPAS. For example, an ℓ_z cutoff score of -1.47 flagged 40% of the simulated random responders and just 1% of the actual examinees. For ℓ_z = -1.09 the corresponding hit and false positive rates were 62% and 3%, respectively. Table 29 shows some representative ℓ_z cutoff scores, hit rates, and false positive rates for the Static and CAT TAPAS versions.

Figure 3. ROC Curve for Static TAPAS Version (“100% Random” and “Actual” Subgroups).

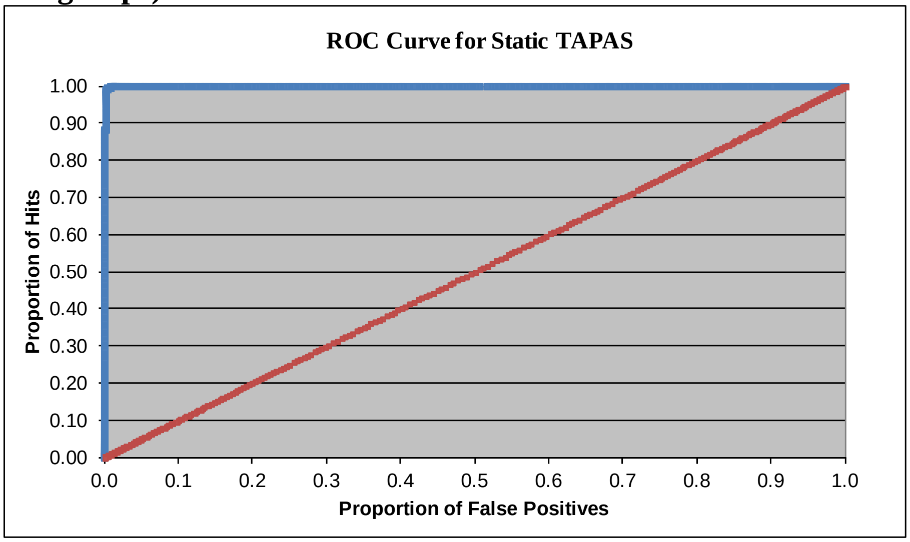

Figure 4. ROC Curve for CAT TAPAS Version (“100% Random” and “Actual” Subgroups).

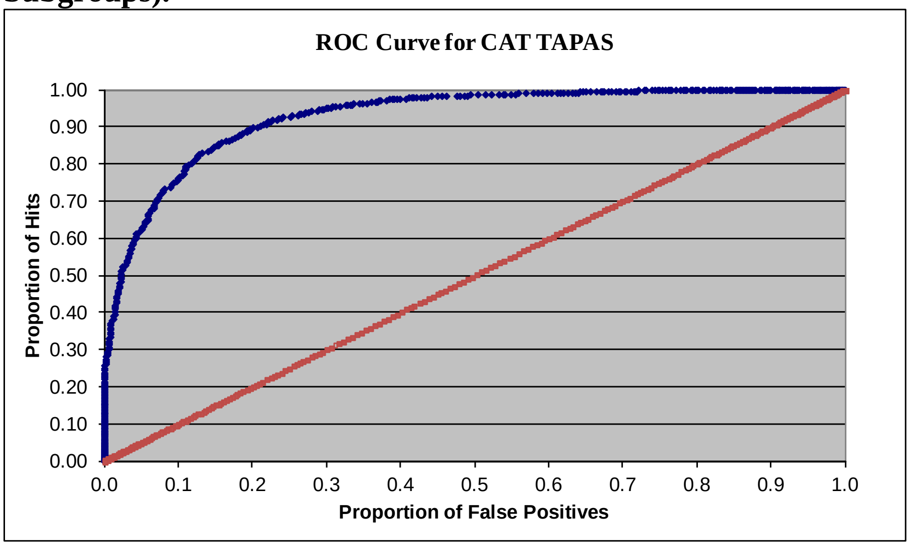

Table 28. Percent of Respondents Having ℓ_z Below the Critical Value

ℓ_z Value	Static		CAT	
	False Positive (Actual)	**Hit (Random)**	**False Positive (Actual)**	**Hit (Random)**
-2.27	0	99	0.1	13.0
-1.47	4.5	99.8	1.0	39.0
-1.23	8.3	99.8	2.0	56.0
-1.09	11	99.8	3.0	62.0
-0.99	13.5	99.9	4.0	66.0
-0.91	16	99.9	5.0	70.0
-0.65	26	100	10.0	79.0
-0.49	23.5	100	15.0	85.0
-0.35	40	100	20.0	88.0
-0.23	47	100	25.0	90.0
-0.13	52	100	30.0	93.0

Summary and Conclusions

In sum, our results suggest that ℓ_z can be used to identify examinees who provide completely random responses. For the 120-item Static TAPAS, an ℓ_z critical value of -2.27 identified 99% of simulated random responders without flagging a single real examinee. For the 120-item CAT TAPAS, about 40% of simulated random responders were correctly identified, along with just 1% of actual examinees; and altering the cutoff score slightly identified 62% of the simulated random responders and just 3% of the real examinees. Thus, ℓ_z may provide an effective screening method for this extreme form of aberrance.

In anticipation of using ℓ_z for screening with the upcoming CAT versions (Versions 9, 10, and 11), we also simulated 1,000 completely random response patterns for each test in an effort to establish ℓ_z critical values that can be used for real time decision making. To flag at least 40% of truly random responders, we found that an ℓ_z critical value of -1.98 would be appropriate. Consequently, that value has been programmed into the new EXE that has been delivered to DMDC for deployment.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

This report summarizes research that was conducted to examine potential moderators of TAPAS test score validities. The research proceeded in three phases. In phase 1, the TAPAS scoring algorithm was generalized to allow for non-standard normal and beta prior distributions that could serve as alternatives to the currently used multivariate standard normal in an effort to thwart random and strategic responding. In phase 2, Drasgow et al.'s (1985) ℓ_z index was adapted for use with MDPP tests and its performance was evaluated with nonadaptive and adaptive tests, similar to those used in the MEPS, via Monte Carlo simulation. In addition, a Markov chain method was developed for identifying patterned responding and flags based on item response times were implemented via SPSS syntax. In Phase 3, the response pattern flags were applied to data collected from a large sample of Army applicants to assess the extent to which TAPAS test score validities are lower for potentially aberrant groups of responders.

From the analyses described in the previous chapters, our key results and recommendations are:

1) Alternatives to the N(0,1) prior distribution that is in the current TAPAS software appear unlikely to have a marked effect on TAPAS criterion validities. However, using prior distributions with lower means could reduce mean test scores for random and/or strategic responders, relative to conscientious responders, and thus influence selection decisions.

2) The MDPP ℓ_z method for detecting aberrance using person-specific critical values provided better than expected power for detecting random and strategic responding with CAT and very high power for detecting aberrant responding with nonadaptive tests, even at strict alpha levels. In this research, we focused on tests involving 120 items and 15 dimensions to emulate the testing configurations in the MEPS. However, the methodology could easily be applied to alternative test configurations and implemented in the TAPAS software to allow real time reporting of ℓ_z statistics (along with Markov chain and response time flags) in conjunction with TAPAS scale scores and Army Can Do and Will Do composites. Future research should examine the extent to which model-data misfit affects MDPP ℓ_z Type I error and power rates to determine whether an alternative way of computing ℓ_z critical values might be beneficial for applications where some model misfit is expected.

3) The Markov chain method provides an easy way of detecting patterned responding and could ultimately be incorporated into the TAPAS software along with ℓ_z to provide an additional real-time screen of examinee responses. Like flags based on item response latencies, the advantage of this approach is that it is simple to compute and makes no assumptions about model fit.

4) Despite the relatively small differences in criterion validities that were observed in the cleaned and full samples in Chapter 4, we recommend screening response patterns prior to validity analyses using a combination of the flags developed here. This recommendation is based on the fact that correlations of some TAPAS scales were lower for the flagged samples than for the clean sample.

5) We believe that the response time screen is the most important flag for detecting aberrance because of the large differences in correlations of TAPAS scales with criterion variables observed for this group relative to the clean sample. With ℓ_z we suggest using a very small alpha level, corresponding perhaps to the 1st or 2nd percentile, to reduce the likelihood of Type I errors because our analyses suggested that, even with person-specific critical values, ℓ_z might be sensitive to departures from model predictions. If that is the case, for example, answering carelessly on a subset of items in an effort to get through with an exam and on to the next task in the MEPS setting could result in a false positive error (aberrance detected), even though an examinee provided relatively "good" data and was assigned trait scores that were fairly reflective of his/her true personality characteristics.

REFERENCES

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumption of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19,* 88-106.

Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical indices. *Applied Psychological Measurement*, *11,* 59-79.

Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1991). Multi-test extensions of practical and optimal appropriateness indices. *Applied Psychological Measurement, 15,* 171-191.

Drasgow, F., Levine, M.V. & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Selection and Classification Decisions* (Technical Report 1311). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Goldberg, L. R. (1990). An alternative "description of personality: " The Big Five factor structure. *Journal of Personality & Social Psychology, 59*, 1216-1229.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277-298.

Knapp, D. J., & Heffner, T. S. (Eds.). (2010). *Expanded Enlistment Eligibility Metrics (EEEM): Recommendations on a non-cognitive screen for new soldier selection* (Technical Report 1267). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Knapp, D. J., Heffner, T. S., & White L. (Eds.) (2011). *Tier One Performance Screen initial operational test and evaluation: Early results* (Technical Report 1283). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Levine, M.V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, *35,* 42-56.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-289.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321-336.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person-fit indices. *Psychometrika, 55,* 75-106.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 213-229.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied psychological measurement, 21*, 115-127.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general model for unfolding unidimensional polytomous responses using item response theory. *Applied Psychological Measurement, 24,* 2-32.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*, 41-54.

Snijders, T.A.B. (2001). Asymptotic distribution of person fit statistics with estimated person parameters. *Psychometrika*, *66*, 331-342.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment* [Doctoral Dissertation]. University of Illinois at Urbana-Champaign.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29,* 184-203.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15*, 463-487.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. (2006). Investigating the appropriateness of an ideal point response process for personality data. *Journal of Applied Psychology*, *91*, 25-39.

Stark, S., Chernyshenko, O.S., Drasgow, F., & White, L.A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15*, 463-487.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273-286.

van Krimpen-Stroop, E. M., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23,* 327-345.